Sampling Strategies for the Proposed National Children's Study

October 25, 2002

Prepared for:

National Center for Health
Statistics
Hyattsville, Maryland
Contract # HHS-100-97-0017

Prepared by:

WESTAT
1650 Research Blvd.
Rockville, Maryland 20850

TABLE OF CONTENTS

Chapter Page

1 INTRODUCTION 1-1

2 THE THREE SAMPLING MODELS 2-1

2.1 Choice of Model 2-2
2.2 The Household Model 2-3

2.2.1 Age Range 2-7
2.2.2 Choice of Respondent 2-8
2.2.3 Under-Reporting of Pregnancy 2-9
2.2.4 Response Rates 2-12
2.2.5 Gestational Age at Enrollment 2-13
2.2.6 Live Birth Rates 2-14
2.2.7 Mechanics of Sampling and Recruiting 2-15

2.3 The Office Model 2-18

2.3.1 Coverage 2-18
2.3.2 Office Selection 2-19
2.3.3 Office Operations 2-20
2.3.4 Recruitment of Pregnant Women 2-20
2.3.5 Response Rates 2-21
2.3.6 Gestational Age at Enrollment 2-22
2.3.7 Office Timeline 2-23

2.4 The Center Model 2-25

2.4.1 Selection of the Centers 2-25
2.4.2 Recruitment of Pregnant Women 2-25
2.4.3 Collection of Baseline Questionnaire Data 2-26
2.4.4 Gestational Age at Enrollment 2-26
2.4.5 Center Timeline 2-27

2.5 Collection of Pre-Pregnancy Data 2-28
2.6 Oversampling Rare Demographic Domains and High-Risk Populations 2-29
2.7 Comparisons of the Models 2-30

3 ALTERNATIVE SAMPLE DESIGNS 3-1

3.1 Household Model Sample Designs 3-2
3.2 Office Model Sample Designs 3-8
3.3 Center Model Sample Designs 3-10
3.4 Sample Designs Considered 3-13

TABLE OF CONTENTS (CONTINUED)

Chapter Page

4 POWER PROJECTIONS 4-1

4.1 Analysis Methods 4-1
4-2 Theory for Power Calculations for Comparisons of Exposed and Unexposed Subjects 4-2
4.3 Assumptions for Power Projections 4-4
4.2 Statistical Power Tables 4-7

5 BIAS AND OPERATIONAL CONSIDERATIONS 5-1

5.1 Bias Considerations 5-1

5.1.1 Undercoverage 5-2
5.1.2 Nonresponse 5-3
5.1.3 Representativeness 5-6

5.2 Data Collection Issues 5-7

5.2.1 Control of Survey Operations and Consistency of Measurements 5-7
5.2.2 Short-Term Considerations for Human Measurements 5-8
5.2.3 Short-Term Considerations for Physical Environmental Measurements 5-9
5.2.4 Collection of Tissues in the Delivery Room 5-10
5.2.5 Long-Term Considerations 5-11

5.3 Hybrid Designs and Organization Structures 5-13

6 COST PROJECTIONS FOR SAMPLE RECRUITMENT AND BASELINE INTERVIEWS 6-1

6.1 Components of Cost for the Household Model Designs 6-2
6.2 Components of Cost for the Office Model Design 6-4
6.3 Components of Cost for the Center Model Design 6-6
6.4 Integration of Data Management Systems 6-8
6.5 Cost Estimates for Recruitment and Baseline Interviews 6-8

7 OVERALL SUMMARY 7-1

REFERENCES R-1

List of Tables

Table Page

2-1 Preliminary estimates of sample sizes for Household Model 2-6

TABLE OF CONTENTS (CONTINUED)

List of Tables (continued)

Table		Page
2-2	Live births by age, United States, 2000	2-8
2-3	Week at which women claimed to learn of their pregnancies	2-14
2-4	Live birth rates by gestational age at enrollment	2-15
2-5	Week at which pregnant women first seek prenatal care	2-22
2-6	Weeks of gestation at time of registration from the Collaborative Perinatal Project (1/1/1959-12/31/1965)	2-27
3-1	Design effect contributions at the PSU level	3-5
3-2	Design effect contributions at segment level	3-6
3-3	Design effect contribution from sampling offices	3-10
3-4	Design effect contributions from centers	3-11
3-5	Critical value for hypothesis tests given various degrees of freedom	3-13
3-6	Possible designs for the National Children's Study	3-13
4-1	Effective sample sizes for an exposed crossclass of 15,000 youth and an unexposed crossclass of 60,000 youth for the four illustrative designs	4-6
4-2	Effective sample sizes for an exposed segregated class of 15,000 youth and an unexposed segregated class of 60,000 youth for the four illustrative designs	4-6
4-3	Projected power for the Household Model A with 800 PSUs and 12,500 segments: Crossclasses of 15,000 exposed and 60,000 unexposed youth	4-10
4-4	Projected power for the Household Model B with 300 PSUs and 3,125 segments: Crossclasses of 15,000 exposed and 60,000 unexposed youth	4-10
4-5	Projected power for the Office Model with 800 PSUs and 4,000 offices: Crossclasses of 15,000 exposed and 60,000 unexposed youth	4-11
4-6	Projected power for the Center Model with 100 centers: Crossclasses of 15,000 exposed and 60,000 unexposed youth	4-11

TABLE OF CONTENTS (CONTINUED)

List of Tables (Continued)

Table Page

4-7 Projected power for the Household Model A with 800 PSUs and 12,500 segments: Segregated classes of 15,000 exposed and 60,000 unexposed youth 4-12

4-8 Projected power for the Household Model B with 300 PSUs and 3,125 segments: Segregated classes of 15,000 exposed and 60,000 unexposed youth 4-12

4-9 Projected power for the Office Model with 800 PSUs and 4, 000 offices: Segregated classes of 15,000 exposed and 60,000 unexposed youth 4-13

4-10 Projected power for the Center Model with 100 centers: Segregated classes of 15,000 exposed and 60,000 unexposed youth 4-13

6-1 Costs for recruitment and baseline interviews for each of the four designs 6-9

List of Figures

Figure Page

2-1 Timeline for Household Model.......... 2-17

2-2 Timeline for Office Model.......... 2-24

2-3 Timeline for Center Model 2-32

3-1 Clustering Decisions for Household Model Sampling 3-3

3-2 Clustering Decisions for Office Model Sampling.......... 3-9

1. INTRODUCTION

The Children's Health Act of 2000 (PL 106-310) mandated the National Institute of Child Health and Human Development (NICHD), the Environmental Protection Agency (EPA), and the Centers for Disease Control and Prevention (CDC), to plan, develop, and conduct a prospective cohort study, now known as the National Children's Study (NCS). The goals of this study are to address important medical, environmental, and social hypotheses concerning the effects of environmental exposures on children's physiology, emotional development, and cognitive abilities. The NCS will be national in scope and will involve the study of approximately 100,000 children from a point in time before birth through age 21. The sample size of 100,000 is just a working number for current purposes. It may be revised later as the plans for the NCS develop. The NCS data collections will be extremely complex, involving the collection of biologic data,[1] environmental exposure data, and neighborhood data, as well as personal information reported by mothers and children. The study's findings should be generalizable as closely as possible to the population of the United States as a whole. The study should also include sufficient sample sizes of children in a sizable number of groups of particular interest to produce separate reliable estimates for these groups.

Under contract to the National Center for Health Statistics (NCHS), a constituent component of CDC, Westat has examined a number of candidate sample frames and sample designs for NCS enrollment. The purpose of this report is to describe and evaluate the various designs, with particular emphasis on four that illustrate a reasonable range of possibilities for the NCS. In proposing and evaluating alternative sample designs, attention was given to questions of generalizability, precision, bias, operational features, and cost. At this time, it is impossible to assess total cost, since it is dependent on the data collection protocols throughout the life of the survey, and these have not yet been established. Westat was, therefore, tasked with developing cost estimates only for drawing the sample, developing a baseline questionnaire for pregnant women, programming a data management system, recruiting the sample of pregnant women for the NCS, and conducting the baseline interview.

As a rule, large-scale national surveys employ probability sample designs so that the survey results will reflect data for the total population using statistical methods that do not depend on untestable assumptions. The starting point for a probability sample design is to identify a sampling frame from which the sample can be selected. Given a frame, the next step is to develop an efficient sample design

[1] In planning the sample designs for this report, the biologic data were taken to be samples of hair, blood, and urine. Members of an expert panel that discussed a draft of this report emphasized the importance of collecting placentas and cord blood samples. This emphasis came too late to guide our research, but some note has been made about how this emphasis affects the alternative sample designs discussed here.

for selecting the sample from that frame. This report considers two alternative frames and associated models for sample selection for enrolling expectant mothers into the NCS. Since there are concerns about the feasibility and costs involved in selecting a national probability sample of pregnant women for the NCS, the report also considers a nonprobability approach and associated model for this purpose. Initially a sizable number of alternative sample designs were considered under each of the three models. Based on the initial review, four of the sample designs were chosen for more detailed study.

The three models for the recruitment of a sample of pregnant women for the NCS are discussed fully in Chapter 2. Briefly, they are as follows:

- **Household Model.** This model employs an area sampling frame, as is generally used for household surveys that involve face-to-face interviewing (random digit dialing telephone interviewing was ruled out as an effective way to recruit pregnant women for the NCS). The model employs home-based screening, in which a large national probability sample of households is contacted and screened to locate women of childbearing ages. The selected women who are not surgically sterile are followed for three years and are enrolled for the NCS if they become pregnant. To yield the desired number of 100,000 births for the NCS, this model requires screening a sample of 1.1 million households to find age-eligible women, and the participation of about 500,000 such women in the 3-year followup study. The sample of age-eligible women is selected in four stages, with probability sampling at each stage: metropolitan areas or counties are selected as the primary sampling units (PSUs) at the first stage; segments (combinations of Census blocks) are selected in selected counties at the second stage; households are selected within selected segments at the third stage; and all age-eligible women in selected households are selected at the fourth stage. Multiple births and births from multiple pregnancies by sample women would all be selected.

- **Office Model.** This model uses a list of physicians on the master files of the American Medical Association (AMA) as a **sampling** frame. A sample of pregnant women who obtain prenatal care from a sample of physicians identified from the files as possibly providing prenatal care is enrolled in the NCS. Again, the sample is selected in four stages using probability sampling at each stage: first, a sample of PSUs (metropolitan areas or counties) is selected; second, a sample of physicians is selected in the selected PSUs; third, for each selected physician, a sample of one of the offices where he or she works is selected; and, fourth, all the women seeing the physician for prenatal care for the first time in a given period are selected. Multiple births from the same pregnancy would all be sampled, but multiple pregnancies would generally not be selected from the same woman.

- **Center Model.** This model employs a nonprobability approach. It involves the selection and funding of a number of health care institutions, termed "centers" hereafter, through a competitive **procurement** process. Each center—which can be thought of as a PSU—is responsible for recruiting a number of pregnant women over a 3-5 year time period. It is envisaged that the centers would be responsible for the NCS data collections throughout the life of the study. A coordinating center would control the standardization of the data collections. The sample design can be viewed as a two-stage sample, with centers selected at the first stage and pregnant women at the second stage. Multiple pregnancies by the same woman might or might not be selected.

Each of these models assumes that the sample of pregnant women is selected in more than one stage—four stages for the household and office models, two stages for the center model. The use of such multi-stage sampling results in a clustering of the sample in the selected PSUs, second stage units (segments for the household model, physicians for the office model), etc. This clustering substantially reduces survey costs and aids operational feasibility. However, it lowers the precision of survey estimates for a given sample size: the greater the clustering of the sample in a small number of PSUs, second stage units, etc., the greater the loss of precision. An important consideration in sample design, therefore, is the degree of clustering of the sample, that is, the numbers of PSUs, second stage units, etc., selected. The effects of different degrees of clustering are examined in Chapter 3, which considers a range of different sample designs for each of the three models. That chapter also identifies the four designs for which cost and power estimates were prepared.

To be effective, sample design and data collection must be considered together in developing an efficient overall survey design that can be implemented in practice. This was done to the maximum extent feasible for each of the three models in developing the sample designs outlined in Chapter 3. However, since the NCS measurement protocols for health outcomes and environmental exposures are still at a very early stage of development, the sample designs outlined in Chapter 3 were constructed with only some general ideas of possible data collection procedures. Consequently, the decisions on clustering made for these sample designs should be viewed as a first iteration, to be reassessed as the data collection procedures are firmed up.

Chapter 4 presents some illustrative power projections for the four designs identified for further study at the end of Chapter 3. The projections focus on the ability of each design to detect an effect of a given exposure, say, on the risk of a rare health outcome.

Chapter 5 assesses and compares various risks of bias associated with the alternative designs. It also addresses some of both the short- and the long-term operational considerations involved in collecting health and environmental exposure data with the alternative designs. The possible role of “hybrid” designs is discussed as well as more general considerations of organizational structures under the alternative designs.

Chapter 6 contains cost projections for drawing the sample, developing a baseline questionnaire, programming a data management system, enrolling the sample of pregnant women, and conducting baseline interviews. These projections are provided for each of the four designs chosen for detailed study. Given the very different methodologies associated with the different designs, careful attention needs to be paid to the assumptions underpinning the cost projections. The chapter therefore contains a detailed account of these assumptions. Furthermore, it should be noted that the procedure for conducting the baseline interview, and hence its cost, is affected by the methods that will be used to collect the environmental exposure, biologic, and neighborhood data. The cost projections will therefore need to be revisited as plans for the NCS data collection develop.

Chapter 7 presents an overall summary of the findings from this project.

2. THE THREE SAMPLING MODELS

This project was initiated with very few constraints on the options that could be considered for sampling pregnant women for the NCS. It was specified that there should be 100,000 new born infants who are examined soon after birth, that their mothers should be enrolled as close to conception as possible (even pre-pregnancy if feasible), that the children would be followed for 20 years to assess various health outcomes, including both medical tests and reporting by the children and parents, and that environmental assessments would be made within respondents' homes and neighborhoods. No information is yet available on the environmental assessments and health outcome measurements. The project required a detailed review of four sample design options that would be chosen on grounds of feasibility, being illustrative of the range of feasible designs, and having favorable properties with respect to either bias, variance, cost, or collecting complex measurements.

The starting point for the project was the identification of the following three basic approaches which seem capable of providing the required number of pregnant women reasonably early in their pregnancy:

1. Screening a large sample of households to identify and locate pregnant women;
2. Selecting a large sample of physicians and medical offices and securing their cooperation to recruit pregnant women seen in their practices; and
3. Selecting and funding a small number of large health care centers to recruit volunteer samples of pregnant women.

These three approaches together with the general operational procedures assumed for them are referred to as the Household Model, the Office Model, and the Center Model. For each model, variations in the sample selection methods can be considered. This chapter describes the three models and their properties. Chapter 3 considers alternative sample designs for each.

An important assumption underlying the choice of these models is that it is essential to begin health and related measurements during pregnancy. Most of the cost of sample selection would be eliminated and a more efficient sample could be selected if the first measurements were delayed to approximately 6 to 12 months after birth. This could be accomplished by selecting, whenever possible, a sample of births reported in the Birth Registration System. (This sampling method is used in the National Center for Education Statistic's Early Childhood Longitudinal Study—Birth Cohort, commonly referred to as ECLS-B.) This frame would provide better coverage, and probably improved response rates, than the designs studied here. However, the current plans for the NCS require health measurements and the collection of certain other data to be made as early in pregnancy as possible. A birth registration sample was therefore ruled out.

Another sample design that had to be ruled out was one in which women are sampled in delivery rooms. This design would involve a probability sample of hospitals, birthing centers, etc., and a probability sample of women giving birth in the sampled locations. This type of design is an attractive one for the collection of placentas and cord blood samples. It was, however, excluded from consideration because it fails to provide the required prenatal data.

Going in the opposite direction, rather than deferring the sample selection until after birth, the U.S. Government agencies with principal responsibility for planning the NCS have indicated an interest in obtaining data on pre-pregnancy health and environmental conditions for the women whose children comprise the NCS sample. We have not included this feature in our analyses of the various sample designs. However, some observations on the ability of the various models to accommodate this feature are given at the end of this chapter.

2.1 Choice of Model

In choosing a sample design for the NCS, a critical decision is whether probability sampling can be used or whether it is necessary to resort to purposive sampling. The advantage of probability sampling is that it provides assurance that the selected children mirror the characteristics of the U.S. population of children, including appropriate representation of small population groups, such as minorities, rural children, etc.[2] As a result, all the survey estimates can be shown to be valid estimates for the population of U.S. children from which the sample was drawn without strong recourse to statistical assumptions.[3] This is particularly important for a “descriptive” estimate of the proportion of the population who have been exposed to a given risk factor, but it also important for an “analytic” estimate of a population exposure-outcome association unless it can be confidently assured that the association does not vary across known or unknown subgroups. In addition, the sampling variance of all estimates can be measured. Purposive sampling lacks these important features. Instead, inferences based on such

[2] It should be noted that probability sampling is not restricted to sampling with equal probabilities. It allows, for example, for certain small subgroups of particular interest to be sampled at higher rates in order to yield adequate subgroup sample sizes to permit separate subgroup analyses. Weights are used in the analyses of the survey data to compensate for unequal selection probabilities.

[3] In practice, probability samples suffer imperfections from noncoverage and nonresponse. Some assumptions are necessary to compensate for these imperfections. However, provided that the extent of these imperfections is not great, the reliance on assumptions with probability samples is modest in comparison with the strong dependence that is necessary in drawing conclusions from purposive samples.

samples rely strongly on models (such as models of disease distributions) as a basis for claims of validity. Where operationally and economically feasible, a probability sample design is therefore preferred for large-scale surveys, and probability sampling is used routinely by government statistical agencies around the world under these conditions.

The Household and the Office Models have been developed as bases for selecting probability samples of pregnant women whereas the Center Model employs purposive sampling. The choice between the three models depends on costs, feasibility, and various other factors associated with the implementation of the alternative models. These factors are identified in the following sections, which describe the models and their properties in detail.

2.2 The Household Model

This section describes the main features of the Household Model, including the number of households that would have to be screened, age limits for screening, problems related to achieving adequate coverage of pregnant women, response rate issues, methods of clustering sample households to keep travel costs at a reasonable level, the mechanics of the sample draw, and similar issues that need to be considered in evaluating the alternative models. These features of the Household Model are identical for all the household sample designs studied for this report.

The Household Model is based on well-established methods used for selecting national probability samples of U.S. households and persons. In general, there is a choice between two main sampling frames for such samples. If the survey data can be collected by telephone, some form of random digit dialing (RDD) sampling approach can be employed using 10-digit telephone numbers as the sampling frame. If the data are to be collected by face-to-face interviewing in the home, then an area sampling frame is used. An area sampling approach has been selected for the Household Model because RDD sampling would not cover households without telephones, and further, because telephone recruitment of pregnant women for the NCS would be inadequate, largely because of problems of response rates in RDD surveys, problems of failing to report pregnancies in a telephone interview, and difficulties of integrating a telephone screening with other aspects of the data collection.

An area probability sample design is selected in several stages. At the first stage, a number of large areas such as metropolitan areas or counties are selected. These first stage areas are known as primary sampling units (PSUs). In each selected PSU, block statistics from the previous Decennial Census are used to combine adjacent blocks into clusters of blocks that are of adequate size for the purposes of the sample. These clusters of blocks are termed segments in this report. The second sampling stage consists of drawing a sample of these segments in each selected PSU. The selection of segments is followed by an operation which lists the housing units in each selected segment. The third sampling stage consists of drawing a sample of the listed housing units. Each sampled housing unit is then contacted and, if occupied, all household members eligible for the survey are listed. The final stage of sampling consists of selecting the requisite number of eligible household members for the survey—often either all or only one, depending on whether it is considered acceptable to interview more than one person in a household.

This area sampling approach could be applied as described to select the pregnant women for the NCS. However, pregnant women are a rare population, and in consequence a very large number of households would need to be screened to give the large sample size required. The Household Model therefore employs a modification that involves drawing a smaller sample of households, identifying women of childbearing age, and following those women who have not been surgically sterilized for an extended period of three years.[4] Sampled women who are pregnant at the time of the first interview or who become pregnant in the followup period are selected for the NCS. The followup procedures ask the sampled women to call the organization conducting the screening as soon as they become aware that they are pregnant, as well as requiring the organization to recontact all sample women at regular intervals to determine their pregnancy status. Where possible, the recontacts would be made by telephone, but some recontacts would need to be made by personal visits to the households. By spacing the recontacts at 3-month intervals, it will be possible to identify pregnancies at an early stage, as desired for the NCS. Less frequent visits, say at 6 or 9 month intervals, would be less costly but not as effective in identifying early pregnancies. This system for finding about new pregnancies is subsequently referred to as the "pregnancy monitoring system."

The initial screening contact with age-eligible women could be used to select all those currently pregnant for the NCS, or it could include only those in the early stages of their pregnancies. For the Household Model discussed in this report, the decision was made to include all those pregnant at the first contact. However, a decision to exclude those in their later stages of pregnancy could be readily accommodated by a slight increase in the initial sample, a slight extension in the period of followup, or a combination of the two.[5]

[4] Note that this period could be easily extended if yields fell below expectations.

[5] Members of an expert panel who reviewed a draft of this report recommended that those found in the initial sample who are past their first trimester should probably not be included in the study. At the same time, it was noted that the recontacts will include some women later in pregnancy who were unwilling to reveal their pregnancy at an earlier time. It was recommended that these late reporters be retained in the sample.

At the household level, it is assumed that the plan will be to recruit all age eligible women (15 through 44 years of age for this study). In most households all of the age-eligible women will be adults but, obviously, many households contain teenage girls and their mothers. The Household Model assumes that in this situation, both the mother and the daughter will be selected for recruitment. Furthermore, if there are multiple teenage girls, it is assumed that all will be selected for recruitment. If this is determined to be too awkward, subsampling can be used, but this would increase both costs and variances. Subsampling within households would require screening more households to reach the goal of 100,000 children in the sample.

Another consideration is that some of the sampled eligible women will have more than one pregnancy during the followup period. To produce a representative sample, all pregnancies and all births (including multiple births) will be included in the sample. From the sampling perspective, this feature leads to an extra stage of clustering. The extra stage produces no analytic difficulties. With appropriate methods of variance estimation for complex sample designs, this effect is automatically taken into account. On the positive side, costs of data collection for two or more children in the same household will be lower than those for two or more children in different households, and indeed it may be possible to use a single measurement of some environmental variables for all sampled children in the same household. Also, the selection of two or more children in some households will permit some sibling analyses (although the sample size will be small).

Before presenting details on sample sizes, one final point should be noted: Not all pregnancies result in live births. Hence, the sample of pregnant women needs to be somewhat larger than the 100,000 live births that has been specified for the NCS (see Section 2.2.6 below).

Other than the long-form sample of the Decennial Census or its planned replacement, the American Community Survey, the Household Model for the NCS would involve screening the largest sample of the American public ever conducted. Table 2-1 presents some basic numbers on sample sizes using this model. Based on current demographic data and experience in other large population surveys, we estimate that a sample of 1.3 million listed dwelling units would be required in order to screen a sample of about 500,000 women in their child-bearing years (ages 15 to 44) for current pregnancy. This estimate is based on an occupancy rate of 88 percent, a household screener response rate of 95 percent, a prevalence rate of 568 age-eligible women per 1,000 households, and a pregnancy screening rate of 80 percent.

Table 2-1. Preliminary estimates of sample sizes for Household Model*

Stage	Sample count
Initial sample of listed dwelling units	1,317,000
Households (residential occupied dwelling units)	1,159,000
Screened households	1,101,000
Age-eligible women and girls	622,000
Age-eligible females screened for current pregnancy	498,000
Number current pregnancies initially reported	16,000
Number additional pregnancies reported over following 35 months	102,000
Total pregnancies reported and baseline interviews given	118,000
Infant exams on live births	100,000

*More work is required to fine-tune these numbers. Research is needed on the coverage that is likely to be achieved of intended and unintended pregnancies and abortion rates.

At the time the sample is initially recruited, some of the women will already be pregnant. When Cycle V of the National Survey of Family Growth (NSFG) was conducted, about 4 percent of women aged 15 to 44 reported being pregnant on the day of the interview. There was probably some underreporting in this NSFG estimate,[6] but many of the unreported pregnancies in the NSFG may have been terminated by abortion so that the underreporting of pregnancies leading to live births was probably slight. It is likely that the NCS will also experience some additional underreporting, especially given the burden in terms of survey response, doctors' visits, tests, and so on that women can anticipate they will be requested to shoulder if they admit to being pregnant. (However, the followup interviews in the NCS panel model may well mitigate against underreporting.) Also, some of the initially sampled women will drop out of the pregnancy-reporting system prior to becoming aware of their pregnancy and even those who report their pregnancy may not consent to the baseline interview. Taking these factors into consideration, we estimate that among those pregnancies where the woman has cooperated with the initial pregnancy screening and does not intend to obtain an abortion, about 80 percent will report on their pregnancies and complete baseline interviews. This projection is based on finer projections of 90 percent reporting and 89 response, as discussed later. With these assumptions and the further assumption that fertility rates will remain similar to those in 1995, we expect that from the 500,000 age-eligible women initially screened for pregnancy, a total of 118,000 pregnancies will be reported where baseline interviews are also granted. Of these, we expect 16,000 to be reported immediately and 102,000 to be reported over the succeeding three years of followup.

[6] See Section 2.2.3.

Assuming that 90 percent of the reported pregnancies with baseline interviews result in live births and that 94 percent of these are examined, that should yield about 100,000 infant exams. These assumptions about underreporting, response rates, and live birth rates are discussed in Sections 2.2.3, 2.2.4 and 2.2.6, respectively.

2.2.1 Age Range

In the Household Model, all women in their child-bearing years in sampled households are asked to participate in the panel screening component of the NCS. A decision therefore needs to be made on the age range for the operational definition of childbearing years. Assuming that the sample selection takes place over a three-year period, the initial minimum age for identification of age-eligible women needs to be lower than the bottom of the age-range. (However, questions on pregnancy do not have to be asked until a sampled female reaches the minimum age.) Table 2-2 shows the number of births by age of the mother. Overall, 99.7 percent of births are to mothers in the age range of 15 to 44 years, with 0.2 percent being to mothers under age 15 and 0.1 percent to mothers over 44. Our workload projections assume that age-eligibility would comprise ages 15-44, with 12-14 year girls identified for later interview. Including mothers aged under 15 in the sample would present severe data collection problems. Including mothers over age 44 in the sample would add to the expense while contributing very few births, but since these children are at higher risk of birth defects, perhaps this decision should be reexamined.[7]

[7] The cost estimates in Chapter 4 assume that women 45 and older will not be recruited.

Table 2-2. Live births by age, United States, 2000

Age of mother at birth of baby	Number	Percent
10	4	0.0
11	16	0.0
12	222	0.0
13	1,427	0.0
14	6,850	0.2
15	21,845	0.5
16	48,581	1.2
17	86,783	2.1
18	132,786	3.3
19	178,995	4.4
20-44	3,576,701	88.1
45	2,281	0.1
46	1,135	0.0
47	527	0.0
48	250	0.0
49	156	0.0
50	100	0.0
51	67	0.0
52	43	0.0
53	30	0.0
54	15	0.0
Total	4,058,814	100.0

Source: National Center for Health Statistics.

Securing informed consent to ask the screening questions of minors is an important concern. In most sample surveys, written consent from adults is required only when physical examinations will be made, physical specimens will be collected, or information will be requested from a third party.[8] However, in the case of minors, greater attention needs to be given to consent conditions for the screening process, including obtaining permission in writing from a parent or guardian.

2.2.2 Choice of Respondent

A major decision to be made with household screening is who is an acceptable respondent for reporting on pregnancy. The least expensive procedure is to ask any adult member of the household

[8] There are, of course, exceptions to this rule. A number of surveys sponsored by the Centers for Disease Control and Prevention have required written consents from adults if the interview covers sensitive topics.

whether any of the women in the household are pregnant, and if no one answers the door, to rely upon information from neighbors. This procedure was tried on the Project on Human Development in Chicago Neighborhoods (PHDCN) with poor results (see Section 2.2.3). A more expensive procedure that results in higher coverage of pregnant women is to ask every age-eligible woman personally about pregnancy, while shielding her replies from other household occupants. This is the procedure used in the NSFG. It was selected for the Household Model.

A rather different sample design is possible in which all women in sample households are asked whether their sisters (and possibly daughters) are pregnant where sisters (and daughters) in other households are in scope. Use of such proxies is called network or multiplicity sampling. By asking women to report not just on their own pregnancies but on those of their sisters, the yield is increased and the cost of the screening is reduced. Further reductions are possible if males are also asked to report on pregnancies of their sisters. Sanders and Kalsbeek (1990) carried out some research on the use of network sampling to identify pregnant women in a small pilot study in a five-county area of central North Carolina. They found that allowing women to report on pregnancies by their sisters doubled the number of reported pregnancies. However, not all of the pregnant sisters could be contacted. The yield of interviewed pregnant women was increased by a net 60 percent. Obviously, if brothers were also allowed to report pregnancies, the boost would be greater, but this was not studied. Network sampling raises three questions. The first concerns the ethics of asking people to report on their siblings. The second concerns coverage. Sanders and Kalsbeek report that only 83 percent of currently pregnant women would tell all their sisters about a pregnancy. This figure drops to 73 percent for unmarried women. It might be that of those pregnancies that are carried full term, the rate at which sisters are informed is near to 100 percent, but this is unknown. Those women who did tell their sisters about their pregnancies reported doing so at about 5.6 weeks after conception, which is encouragingly early. The third concerns the operational feasibility of including sisters so identified in the NCS data collections. Restricting the network to linkages with sisters who live nearby may at least partially address this question, but that would reduce the number of sisters eligible for the NCS. We do not believe network sampling would produce satisfactory coverage, and have not involved it in our list of sample designs.

2.2.3 Under-Reporting of Pregnancy

Achieving a high reporting rate for all pregnancies is an important issue with the Household Model. It will probably depend strongly on the context created by the combination of advance mailings, questionnaire content, and questionnaire administration mode. Also, the panel nature of the recruitment process should help since the women in the panel will be familiar with the survey objectives and, even if they fail to report their pregnancy at one round of contact, they are more likely do so at a later round. In this section, we estimate reporting rates for two prior surveys and then speculate about the rates that might be achieved in the NCS.

As noted earlier, the NSFG Cycle V found 4 percent of age eligible women to be pregnant on the day of interview. Also, from the NSFG, the average gestational age at cognition of pregnancy was 32 days, and the average pregnancy lasted 196 days. Based on these numbers, the window during which a woman could report a pregnancy is 165 days long or 45 percent of a year. With these assumptions, 40 reported pregnancies per 1000 age-eligible women on the day of the NSFG interview would imply a total of 88.5 pregnancies per 1,000 age eligible women over the course of a year. With 60.2 million age eligible women in 1995, that would imply 5.3 million pregnancies per year. Table 85 of the 2001 Statistical Abstract of the United States reports 6.2 million pregnancies for 1995. So probably only about 85 percent of pregnancies were reported to the NSFG interviewers. However, when abortion rates are considered, it seems very likely that many of the unreported pregnancies were later terminated by induced abortion. The same table from the Statistical Abstract shows 1.4 million induced abortions for 1995, a rate of 22 percent. So the fact that about 85 percent of all pregnancies were reported in the NSFG is actually very encouraging given that the NCS has no interest in studying pregnancies that end in abortion.

To understand how such high reporting levels were achieved in the NSFG, it should be noted that Cycle V of the NSFG was administered by a mixture of CAPI and AudioCASI[9] with no proxy response,[10] that all the interviewers were female, and that the question on pregnancy came after many other questions that served to establish the legitimacy of the survey and perhaps to desensitize the respondent (Kelly et al., 1997). An important question is whether such high reporting rates can also be achieved in screening for the NCS.

Research on other surveys (Judkins, et al, 1999, and Horrigan, et al, 1999) indicates that in screening for a rare population group, the targeted group is frequently rarer among the completed screeners than it should be, leading researchers to suspect that respondents sense which answers to the screening questions will excuse them from further participation in the survey process. It also emphasizes the importance of the respondent rules in the screening. We assume that proxy reporting will be

[9] CAPI is a survey administration mode where interviewers read questions from a laptop screen and records respondents' answers. Audio-CASI is a survey administration mode where the respondent listens to the questions on private headphones and then directly enters the answers into a laptop computer herself. Privacy is thus assured.

[10] A proxy response would be an instance where another household member, neighbor, or visiting friend would be allowed to respond on behalf of the sample person.

permissible only as a last resort, since its use could contribute appreciably to underreporting of pregnancy.

As an example of this phenomenon with a focus on the underreporting of pregnancies, consider the experience of the Project on Human Development in Chicago Neighborhoods (PHDCN), a project of the Harvard Medical School. As part of this project, door-to-door screening was used to recruit the initial sample. There was a special interest in sampling pregnant women, but they were not the sole type of eligible respondent. Using answers from whomever answered the door (and sometimes from neighbors), the PHDCN found and interviewed just 79 pregnant women per 10,000 households,[11] appreciably fewer than the NSFG Cycle V figure of about 122 pregnant women per 10,000 households.[12] If we make the assumption that PHDCN and NSFG Cycle V interview response rates were similar and that NSFG had near 100 percent coverage of pregnancies not terminated by abortion, this would imply a coverage for the PHDCN of 65 percent not terminated by abortion.

For the NCS, we projected 90 percent coverage of pregnancies that are not terminated by abortion. This is based on the assumed use of a 10-15 minute pregnancy screening interview, answered by the woman herself. Any shortening of the instrument is likely to increase underreporting. One reason we do not project coverage higher than 90 percent is that current standards for ethical research probably require that informational materials given to the household include mention of the 20 year planned duration of the study, and that this information may lead to some underreporting in addition to nonresponse discussed in the next section.

On the other hand, we assumed higher reporting rates than on the PHDCN because of the exclusive use of self-reporting. Experiments will be required to estimate the reporting rate more precisely. If experiments are not done in advance of data collection or if they are too small too yield reliable estimates, we note that it is possible to shorten or lengthen the three-year followup period.

[11] Based on personal communications with Alisú Schoua-Glusberg of Harvard Medical School. Screening for the baseline of the PHDCN was done in 1996 and 1997. She was unable to identify how many pregnant women there may have been in the screener sample, so it is not possible to separate nonresponse from underreporting.

[12] On Cycle V of the NSFG, the response rate is known to have been about 71 percent, so although the NSFG found about 173 pregnant women per 10,000 households, it interviewed about 122 of the 173.

2.2.4 Response Rates

It is useful to separate response rates for the various phases of the study. In terms of phases, there is the general household screening to find women in their childbearing years. This is followed by the screening of age-eligible women for current pregnancies. Then there is maintaining the panel of eligible women for the duration of the followup period. Next there is having pregnant women answer the baseline questions. After pregnancy reporting, there is the first infant exam. Lastly, comes the long-term followup. First though it is worth noting that there is considerable uncertainty about all the projected response rates. If any of the response rates are lower than projected, it would be easy to adjust the sample by extending the period for asking screened women about pregnancies beyond the recommended three years. Correspondingly, if any response rates are higher than projected, the followup period can be shortened.

For general household screening to find women in their child-bearing years, surveys may be able to resort to the use of information from neighbors after a number of unsuccessful attempts have been made to establish contact with the occupants of a sample household. Assuming that neighbor information is allowed for this purpose, it should be possible to screen residents of 95 percent of occupied housing to identify age-eligible women.

The screening of age-eligible women for current pregnancy by self response will require making several attempts to contact the residents of a fairly large proportion of the sample households. It will also require providing women with information about the full study with its 20 years of followup. We think that this screening will have a success rate comparable to the response rate for an extended interview in other surveys. We are estimating this rate to be about 80 percent. Cycle III of the NSFG in 1982 achieved a response rate of 83.5 percent of eligible women.[13] Somewhat less encouraging, but not far off the mark, response rates for sampled women in screened households varied between 77 and 81 percent for the National Household Survey of Drug Abuse (NHSDA) between 1993 and 1998.[14]

Once women have consented to the initial pregnancy screening, we estimate that among those who become pregnant, intend to carry the baby full term, and are willing to honestly report on their pregnancy, about 89 percent will complete the baseline survey. This sample loss includes loss to followup during the three years of screening for new pregnancies.

[13] See Bachrach, et al, 1985. Cycles IV and V of the NSFG are less relevant because they used subsamples of the NHIS and the long lag between the NHIS interview and the NSFG interview had a severe negative impact on response rates.

[14] See Chromy, et al, 1999. This included women, ages 12 and older.

Estimates of the response rate for the infant exam are more uncertain because the nature of this exam has not yet been determined. Response rates will depend strongly on whether the exam can be conducted in the infants' homes and on how invasive and/or painful the parent believes it will be for the child. For the Collaborative Perinatal Project (CPP), the initial examination rate was 97 percent (Niswander and Gordon, 1972), but that was in a very different setting where the recruiting centers delivered the babies. For Cycle III of the National Health and Nutrition Examination Survey (NHANES), where mothers took their children to mobile examination centers (MECs) or had a home exam, the examination rate for children ages 2 to 11 months where an interview had already been granted was 97 percent for black infants, 95 percent for Hispanic infants and 94 percent for whites and others (Mohadjer, et al, 1996). The Phase 1 examination rate for all sampled children under the age of 6 was 85 percent (Khare, et al, 1994). However, home interviews were granted for 93.9 percent of sampled children so the conditional examination rate was 90.5 percent.

For the same type of exam, the NCS examination rates might be lower than those for NHANES because of the time lag between recruitment of pregnant women and the births of their children. In the NHANES, the lag between recruitment and examination was a few days or weeks. In the Household Model of the NCS, the range of lag times would be much larger. That time lag will mean that a number of women will have moved, some of them long distances. Our estimated infant examination rate of 94 percent may be somewhat optimistic and can probably only be met if the exam places very little burden on the mothers.

The high response rates achieved in NHANES are partially due to cash incentives given to respondents for cooperating. Substantial incentives are likely to be necessary to secure high response rates in NCS, and the provision of incentive and their magnitudes are issues that should be examined during the planning of the physical examination and related measurement system.

2.2.5 Gestational Age at Enrollment

The NCS would like to recruit pregnant women as early in their pregnancies as possible. Of course, a woman needs to know she is pregnant before she can report the pregnancy. Cycle V of the NSFG (conducted in 1995) included a question on how quickly women learned of their pregnancies. The results for the most recent pregnancy within the past five years are shown in Table 2-3. The table includes completed pregnancies that resulted in live births, still births, miscarriages, and ectopic pregnancies. The majority of women claimed to know within a month. Given the standard obstetrical definition of gestational weeks as weeks since the end of the last period, this doesn't seem possible. However, it is likely that many women understood the question to mean weeks since inception. Almost all knew before the end of the first trimester. Advances in early pregnancy testing technology since the early 1990s probably mean that women learn even sooner now. So even if many of the reports under 4 weeks should really be counted as a few weeks later, it should be possible to collect some information on both the mother and the embryo fairly early in the first trimester.

Table 2-3. Week at which women claimed to learn of their pregnancies[15]

Learned of pregnancy at	Percent	Cumulative percent
1 week	3.4	3.4
2 weeks	11.1	14.5
3 weeks	10.4	25.0
4 weeks	24.7	49.6
5 weeks	8.9	58.5
6 weeks	15.8	74.3
7 weeks	3.3	77.5
8 weeks	10.9	88.5
9 weeks	1.2	89.7
10 weeks	1.9	91.5
11 weeks	0.4	92.0
12 weeks	4.7	96.7
13 weeks	0.3	97.0
14+ weeks	3.0	100.0

2.2.6 Live Birth Rates

Table 2-4 shows live birth rates from the CPP by gestational age at enrollment for those women enrolled after the 20th week of gestation. Clearly these data are old (from the early 60s) and are for women who have passed the period of miscarriages and most abortions. Live birth rates are probably higher now for pregnancies that reach 20 weeks of gestation, but miscarriage rates will be an important source of sample loss. To the extent that women would enroll earlier with the Household Model, it is clear that live birth rates will be much lower—due to both induced abortions and miscarriages. However, we assume that women planning abortions will generally not report their pregnancy to the NCS interviewer. So that leaves miscarriages. The earlier that a pregnancy is reported, the more likely it is to result in

[15] The exact wording of the question was, “How many weeks pregnant were you when you learned that you were pregnant [on this pregnancy].” There was further guidance that, “Weeks of pregnancy should be counted as completed weeks since the last normal menstrual period.” However, reading this clarification was optional for NSFG interviewers. It might be that many women understood the question to mean weeks since the actual date of inception.

miscarriage. There are also strong dependencies on the age of the woman (see Cunningham, et al., 2001, page 856). Excluding induced abortions, the percent of pregnancies reported in the NSFG that result in live births is 81 percent.[16] We assumed a live birth rate of 90 percent on the theory that many of the miscarriages will happen early—early enough not to have been reported by members of the NCS screening panel to the pregnancy monitoring system. If, during the course of the study, this turns out to be too low, the enrollment period during which age-eligible women report new pregnancies can be easily extended beyond the recommended three years.

Table 2-4. Live birth rates by gestational age at enrollment

Gestation	Pregnancies*	Live birth rate
1-20	17,987	na
21-24	6,386	98.2%
25-28	5,222	98.2
29-32	4,305	98.6
33-36	3,313	98.9
37-39	1,499	99.2
40+	466	98.5
Unknown	37	81.1
Total	39,215	97.8%

*There were 39,215 women with stillbirths or live births on their first studied pregnancy. The table excludes pregnancies that ended prior to the 20th week of gestation. It also excludes multiple births.

It is worth noting that the miscarriages reported in the Household Model will be analytically interesting and that there will be more of these available for study than in the Office and Center Models since the women will be recruited at earlier gestational ages.

2.2.7 Mechanics of Sampling and Recruiting

The mechanics of sampling for surveys like the one proposed here are well established, but there are suboptions that would need to be determined. A sample of PSUs would be selected with stratification and unequal selection probabilities to minimize between-PSU variances for key items of analysis, subject to the constraint that unbiased estimation of the total variance is possible. It would be

[16] Special tabulation of Cycle V of the NSFG.

possible to include certain hotspots (such as Niagara County, New York), into the sample with certainty or to oversample certain subgroups of particular interest, but the effect of such oversampling has not been examined for this report except for some brief remarks in Section 2.6.

After the PSUs had been drawn, a sample of segments would be selected using information from the Decennial Census on housing counts and locations. A map for every segment would be prepared. All the housing in the segments would need to be listed. We suggest combining the listing and screening into a single operation, as was done for Cycle 1 of the National Survey of American Families (Judkins, et al, 1999).

As the sample is screened, all women of childbearing age in the sampled households would be asked to participate. The first phase of participation would be to provide information on their current pregnancy status and fertility. Those who report that they are pregnant would then be asked to complete the baseline interview.

Those women who are not currently pregnant and not surgically sterile[17] would be provided with information on how to contact the survey organization in the event that they become pregnant. They would also be informed that an interviewer will call back in a few months. Except in the case of "hard refusals," telephone interviewers would call every three months to check for new pregnancies and for first contact with girls who have reached the age for participation in the study. In addition, interviewers would pay repeat personal visits to women who do not have phone service or who do not respond to the telephone calls. Women who move would be followed to their new addresses provided they remain in the 50 states and the District of Columbia.[18] As additional pregnant women are discovered, they would be administered the baseline interview and be fully enrolled in the NCS.

It is estimated that screening 500,000 women for initial pregnancies could be done in about two years, with followup to reach the number of new pregnancies desired taking another three years. The entire schedule for this approach is shown in Figure 2-1, assuming a project kickoff at the first of 2004, following preparatory research in 2003. Allowing for instrument development and piloting, the total elapsed time to complete the recruitment would be 7.5 years.

[17] Chandra and Stephen (1998) report that 27 percent of women aged 15 to 44 in 1995 were surgically sterile. Survey costs are substantially reduced by eliminating call backs to women who report being infertile because of surgery. Even though some of these women might have the surgery reversed and a few will become pregnant despite the tubal ligation, the coverage loss seems small compared to the substantial cost savings.

[18] Following women to the border regions of Mexico and Canada might be considered since many of these might move back and forth across the border.

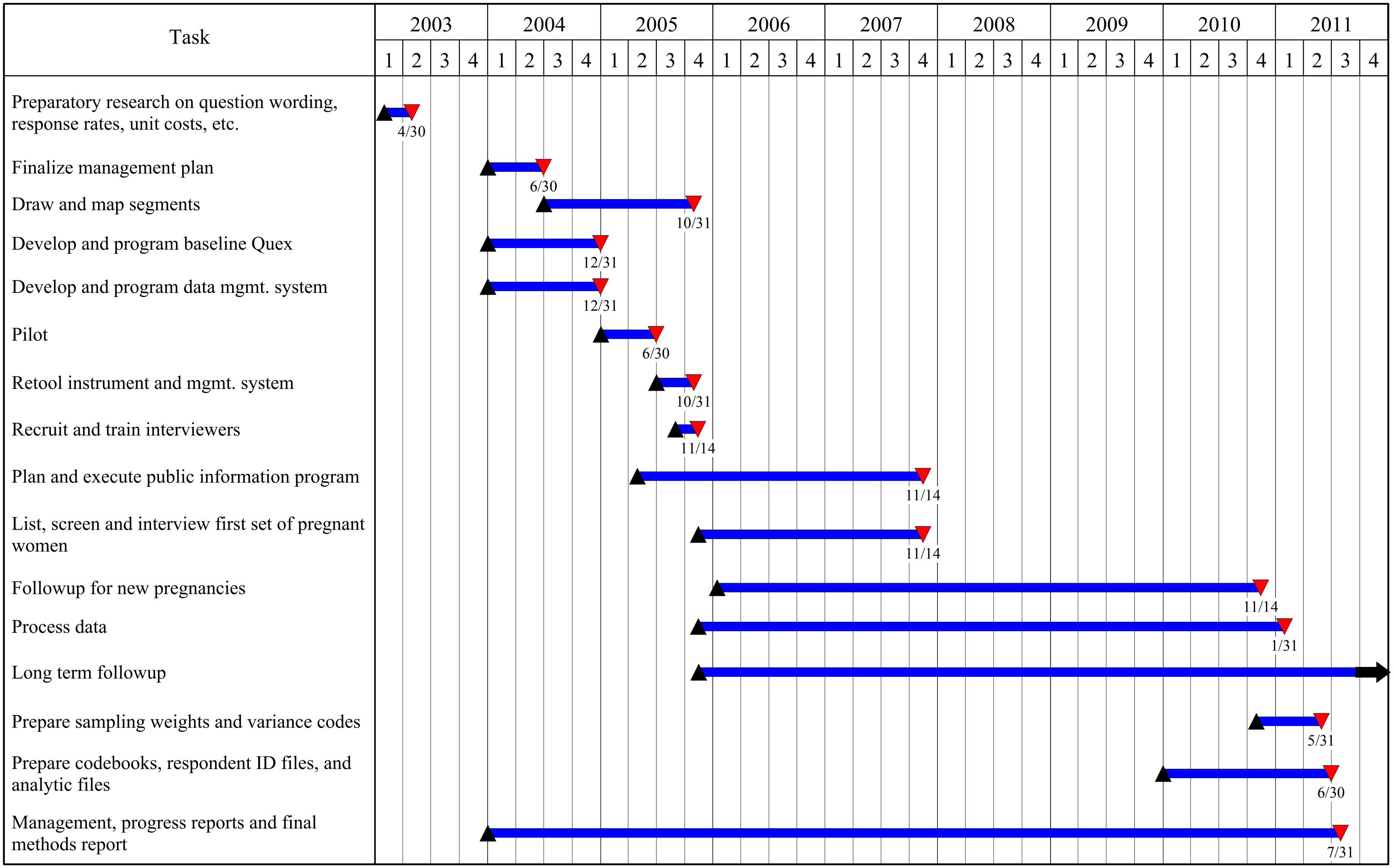

Figure 2-1. Timeline for Household Model

2.3 The Office Model

The Office Model aims to avoid the large-scale screening needed with the Household Model by sampling pregnant women as they visit their physicians for prenatal care. The basic sampling frame is a list of prenatal care providers derived from the master files of the American Medical Association, files that contain listings of virtually all medical doctors (MDs) and doctors of osteopathy (DOS). As with the Household Model, the sample is selected in four stages. First, a sample of PSUs is selected, then a sample of prenatal care providers is selected in sampled PSUs, then one of the offices where the sampled provider works is selected, and finally a sample of pregnant women making their first prenatal care visit at the sampled office is selected. The first stage of sampling PSUs is introduced in order to cluster the sample of pregnant women in selected geographical areas to facilitate NCS fieldwork. The sample of doctors is employed because no list of offices is available; lists of offices are compiled for selected doctors.

This section focuses on the mechanics of the Office Model. It is divided into subsections for coverage, office selection, office recruitment, office operations, gestational age at enrollment, and enrollment timeline.

2.3.1 Coverage

There are two main sources of noncoverage with the Office Model. First, by restricting the sample to women seeking prenatal care, women who do not seek care have no chance of being selected for the NCS. In Cycle V of the NSFG, 2.7 percent of women reported that they did not seek prenatal care (see Table 2-5 in Section 2.3.5). Such women are clearly likely to differ from those who do seek prenatal care in terms of economic status, having health insurance, etc.

Second, the types of doctors selected from the AMA files as potential providers of prenatal care will be incomplete to some extent. Selecting physicians with general practice, family practice, obstetrics-gynecology, obstetrics, gynecology, maternal-fetal medicine, reproductive endocrinology, or infertility as a primary or secondary specialty should capture most physicians providing prenatal care, but some may be missed. Furthermore, physicians who have move their practices between the time of AMA list updating and the time of office recruiting will present a problem. Assuming that the sample will exclude those who have moved away from the PSU, they will be missed. Alternatively expressed, those who move into the PSU will be missed.

Other issues of coverage concern the need to obtain a complete list of a physician's ancillary offices, and the need to ensure that all the pregnant women in the sampled time period are included in the sample. This latter issue is of particular concern because the selection of the sample generally has to be left to personnel in the selected offices.

2.3.2 Office Selection

The sample of physicians would be selected from the lists of physicians likely to provide prenatal care within the sampled PSUs. A two-phase sample would be used. At the first phase a large sample[19] of physicians would be selected with equal probability in each selected PSU.[20] Using information about joint practices from insurance companies, the sampled physicians would be grouped into practices. The office manager of each practice would then be contacted by phone and asked to list all of the practice locations (including those outside of the geographical boundaries of the PSU) for each sample doctor in the joint practice. The office manager would further be asked to roughly estimate the average number of new prenatal patients seen by each doctor at each practice location within some reasonable time period, such as the last six months. Based on this information, a sample of offices (doctors and locations) would be selected using sampling with probability proportional to estimated numbers of pregnant women seen. Given the range of practice models including staff model HMOs, negotiations with some offices are likely to be complex. The interviewers selected for this operation will need to have exceptional talent and training. This has been reflected in the costs presented in Chapter 6.

Kalsbeek and Mancewicz (1993) report that about 21 percent of physicians who provide prenatal care have ancillary practices. This is why it is necessary to inquire about all the locations at which each sample doctor practices, and to give a chance of selection to the pregnant women seen at each location. In some cases, this will include women who reside at some distance from the principal office, including distant counties. These women need to be included since this is the only way they can be selected for the sample. The sampling plan will thus occasionally require extensive travel in the long term followup.

[19] In many PSUs, it might, in fact, be best to select all physicians in the target specialties.

[20] A possible variant is to sample physicians with equal probability within a class of physicians, but with varying probabilities across classes. The purpose would be to sample classes of physicians who are less likely to give prenatal care at lower rates.

The purpose of the subsample with probability proportional to numbers of pregnant women seen is to make the sample more efficient by keeping all the most active offices in the sample and subsampling the least active offices.

2.3.3 Office Operations

Once the offices are selected, the next stage is to secure the cooperation of the selected physicians. This operation would be a challenging and time-consuming one. In addition to obtaining the cooperation of the physician and his or her staff, in a number of cases there will be need to obtain the approval of the Institutional Review Board (IRB) under which the office conducts any research.[21] To encourage participation, doctors and their staff would be compensated in some way such as giving them an incentive for every patient recruited into the sample. One concern that might arise is the long period of intense scrutiny of the children and the possible implications of that scrutiny for malpractice lawsuits. Clearly, participating women cannot be asked to accept any special limitations on their or their children's legal rights. So it will probably be necessary to make sure that the incentives are enough to cover the cost of the extra malpractice insurance associated with each extra patient. Doctors will also need to get informed consent to permit them to provide the patients' names and medical data to the organization carrying out the followup activities and to release medical records. Standard forms should be prepared for the consent. Substitutions would be made for recruitment failures.

2.3.4 Recruitment of Pregnant Women

Each sample physician-practice would be in sample for a fixed number of weeks based upon the workload history. During the target weeks, all women seeking their first prenatal care would be in sample. For simplicity and for better control on the process, there would be no subsampling within the weeks. In those obstetrical offices where the patients are deliberately rotated among the physicians in the practice, we would associate women on the first prenatal care visit with the physician that they happened to see on that visit.

Most of the recruiting work would be done by doctors and their staffs although it might be practical to use professional recruiters in large offices. Whoever does the recruiting would be expected to keep track of both recruitment successes and failures. Additional work would be required to develop and test the best recruitment procedures, but one reasonable model might be to have the doctor briefly

[21] See Wolf, Croughan, and Lo (2002) for a recent discussion of the challenges in this process.

mention to women that either one of his nurses or a professional recruiter will see them next to discuss participation in an important study with them. It might also be sensible to have the recruiting physicians carry out whatever physical examinations are required of the pregnant women.[22]

Although doctors collect their own case histories, it will be important to have pregnant women report their histories as part of the survey's standardized baseline questionnaire. One approach for administering the baseline questionnaire would be to have the women complete the questionnaires on their first visit, using a self-completion paper-and-pencil or computer-based questionnaire. If a computer-based system were used, the NCS could support the physicians and their office staff by supplying them with training manuals, computers, software, information brochures, and a help line. However, the work to train office staff in 4,000 offices seemed too daunting.

An alternative approach would be to have the sampled women give consent to their names and addresses being provided to a survey organization, which would then send interviewers to their homes to conduct the baseline interviews. Telephone appointments could be made to make the personal visits more efficient. The self-completion of the baseline questionnaire in the office is an untried method and its success is uncertain. We have therefore assumed for the Office Model that the alternative approach would be used of conducting baseline interviews by survey interviewers in the sampled women's homes. However, the need for a subsequent visit to obtain the baseline information also will result in some additional refusals.

A concern with the office-based enrollment is the difficulty of controlling the selection and recruitment of the pregnant women. These operations will be conducted without the tight supervisory control that is applied in standard survey settings. Moreover, they would often be left to office staff who lack the training and commitment of professional survey interviewers. Consequently, there is a serious risk that some women who should be sampled will be missed and many others will not be persuaded to participate in the NCS.

2.3.5 Response Rates

We estimate a response rate of about 70 percent in the prescreening of physicians and a 70 percent recruitment rate among those subsampled after prescreening. We have no good idea of the recruitment rate for pregnant women that might be achieved within sampled offices, but a recent study in

[22] Note, however, that the costs presented in Chapter 6 do not include the costs for implementing this suggestion.

Denmark that used procedures similar to those proposed here for the office model obtained a recruitment rate of 60 percent of patients at participating doctors (Olsen, et al, 2001). If we multiply these three together, we get a response rate of 29 percent. We note that the Danish study also obtained a doctor recruitment of 60 percent, somewhat higher than the 50 percent we are estimating.

2.3.6 Gestational Age at Enrollment

Table 2-5 presents a special tabulation of live births reported in Cycle V of the NSFG of the week at which women claim to have first sought prenatal care. Women aged 15 to 44 were asked this question about completed pregnancies that led to live births in the five years preceding the interview where the baby or babies were not surrendered for adoption. As shown in the table, 89 percent of women claimed to have sought prenatal care in the first trimester, 8 percent in the second trimester, less than 1 percent in the third, and 2.7 percent never seek it at all.

Table 2-5. Week at which pregnant women first seek prenatal care[23]

Sought prenatal care at	Percent	Cumulative percent
1 week	0.4	0.4
2 weeks	2.6	3.0
3 weeks	3.3	6.4
4 weeks	13.5	19.8
5 weeks	6.7	26.5
6 weeks	16.4	43.0
7 weeks	4.2	47.2
8 weeks	19.5	66.6
9 weeks	3.0	69.6
10 weeks	5.6	75.1
11 weeks	1.3	76.5
12 weeks	11.6	88.1
13 weeks	0.9	89.0
14-26 weeks	7.8	96.8
27+ weeks	0.5	97.4
Never	2.7	100.0

Source: National Survey of Family Growth, Cycle V.

[23] As noted in connection with Table 2-3, NSFG respondents may frequently have understood weeks to mean weeks since the day of inception rather than the standard weeks since the end of the last menstrual period.

2.3.7 Office Timeline

It is possible that the Office Model could result in a very rapid recruitment of the sample. There are about 31,000 office-based obstetricians and gynecologists in the United States.[24] It is not known what percent of these provide prenatal care nor what percent of other specialists provide prenatal care, but some rough extrapolations from the paper by Kalsbeek and Mancewicz indicate that the total number providing prenatal care is probably fewer than 40,000. Thus a sample of 4,000 would expect to see at least 10 percent of the women who give birth in a year. Since the annual number of births is on the order of 4 million, a sample of 4,000 prenatal care providers should see at least 400,000 new patients in a year. Given the need to recruit 118,000 pregnant women, this indicates that it should be possible to complete the recruitment in as little as 6 months following recruitment of the doctors. Assuming that the doctor recruitment also takes 6 months following six months of prescreening, the entire recruitment period might only last 18 months. A more comfortable timeline might allow 24 months. Such a timeline is shown as Figure 2-2, assuming a project kickoff at the first of 2004, following preparatory research in 2003. Allowing for instrument development and piloting, the total elapsed time to complete the recruitment would be 4.25 years, much quicker than the 7.5 years required with the household model.

The gap between the enrollment dates for the women recruited first and last would be about 17 months. Note, however, that this timeline could be stretched out by dividing the sample PSUs into replicates and staggering the starting dates for the replicates. There are potential analytic advantages in spreading the enrollment over a longer period of time. Spreading the doctor recruitment over a longer time period might lessen some components of cost due to scaling issues, but it would increase the cost of completing the baseline interviews in women's homes because the field force would have to be smaller and travel more when the work is less concentrated in a period of time.

[24] Table 153 of the 2001 statistical abstract of the United States.

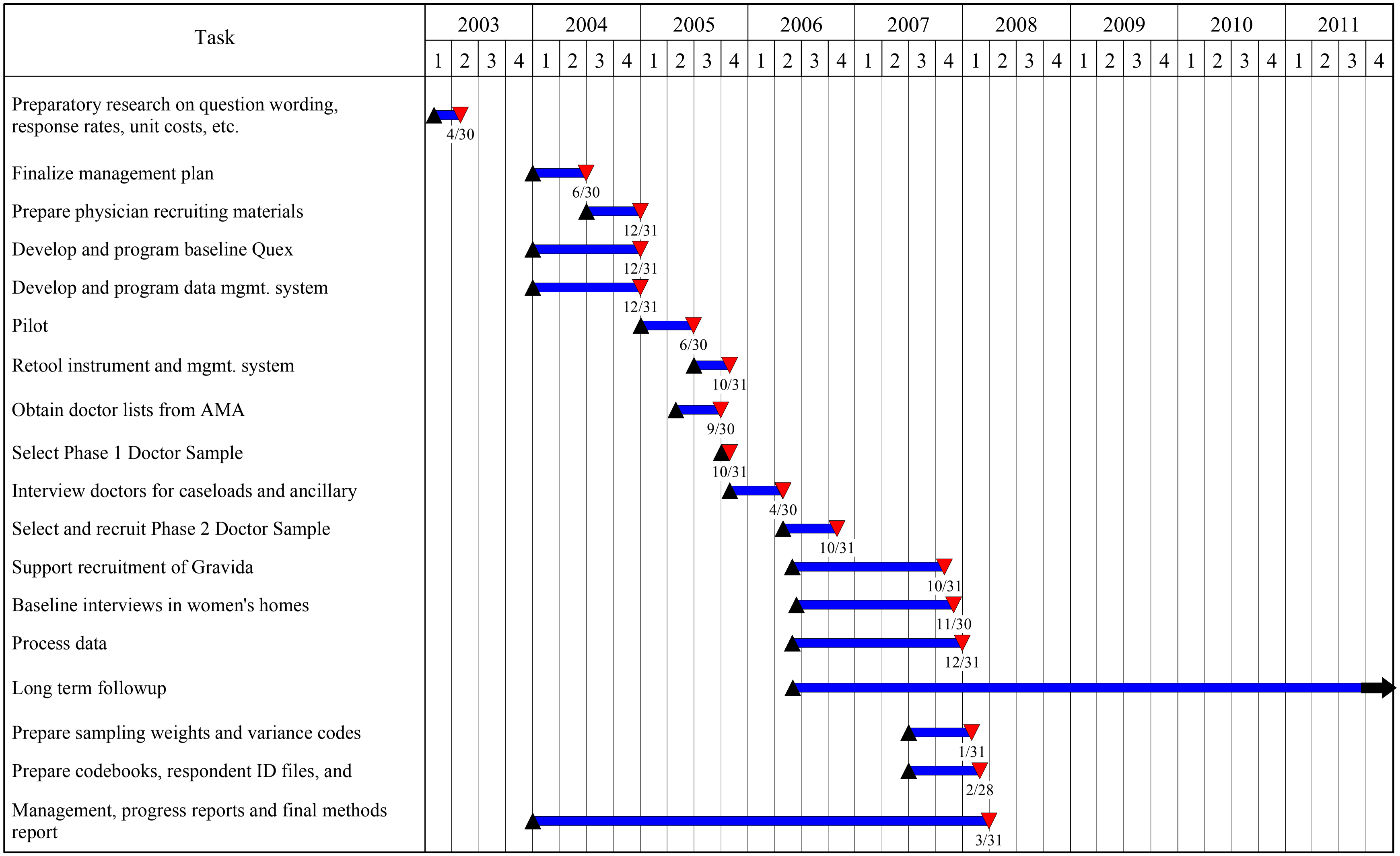

Figure 2-2. Timeline for Office Model

2.4 The Center Model

The Center Model involves the selection and funding of a small number of large health care centers, each of which would be responsible for recruiting the sample of pregnant women. To be manageable, the number of centers would need to be relatively small, but there are also good reasons not have too few: the next chapter considers 100 centers[25]. Based on 100 centers, on average each center would be expected to recruit about 1,200 pregnant women over a three to five year period. The expectation is that the centers would then be responsible for the ongoing measurements and health examinations and for maintaining current records for the children resulting from live births for the duration of the NCS. A coordinating center would be needed to ensure common procedures and comparability of data between the centers. Each center would no doubt have its own IRB, and the coordinating center would need to assure that the standardized procedures satisfy the requirements of all the individual IRBs. Questions of data rights for the center directors might also be a complicating factor.

Areas explored in more detail in the following subsections include: selection of the centers, procedures for recruitment of pregnant women, collection of baseline interview data, expected gestational age at enrollment, and the project timeline.

2.4.1 Selection of the Centers

Some sort of contracting or grant awarding process would be required to establish the network of centers. Bidders might include teaching hospitals, other hospitals, HMOs, rural health clinics, and other clinics. Centers would need to have a charter, space, personnel, and a model for recruiting pregnant women. Some of our assumptions in this area are discussed in Chapter 6. Presumably, there would be goals set for the geographic dispersion of the centers. There might also be goals set for mix of patients with respect to race and socio-economic class.

2.4.2 Recruitment of Pregnant Women

On average, with 100 centers, each center needs to recruit 400 pregnant women per year over a three-year period for a total of 1,200 pregnant women per center. Depending on the nature of a given center, there may or may not be such a strong natural flow of women seeking prenatal care in the

[25] Members of the expert panel suggested that 100 centers is too many and that a number between 20 and 30 would be more feasible.

center to satisfy this requirement. One would expect such a flow at a large staff-model HMO but special efforts might be required at other types of centers. It will be important to plan some external recruitment techniques to supplement the sample before the start of data collection. Some centers may rely on the local mass media (advertisements and public service announcements). Others may use referrals from a network of local doctors. Another possibility is to form a partnership with a set of public health clinics. We understand that many teaching hospitals already have such partnerships in place where interns and residents provide the ambulatory are in the public health clinics.

Whatever the external modes of recruitment, it seems likely that the sample of pregnant women will underrepresent those who do not seek prenatal care. It also seems likely that it will be difficult to find centers with sufficient volume in many rural areas, so that pregnant women in these areas will be underrepresented. This topic is discussed further in Chapter 5.

Under the Center Model, the recruitment of pregnant women could be carried out using some form of quota sampling. Quotas could be set in terms of socio-demographic groups (e.g., by mother's age and race) to try to make the overall sample from all centers combined more closely resemble the larger U. S. population. Alternatively quotas could be used to implement various forms of oversampling as discussed in Section 2.7.

2.4.3 Collection of Baseline Questionnaire Data

In the Center Model, it should be feasible to give the same sort of training on interview procedures to the local staff as is given to professional interviewers in the Household Model. There is thus no need to send interviewers to the home of sample women to conduct the baseline interviews. Instead, it is assumed that Center Staff assist women in completing the baseline questionnaire while visiting the center.

2.4.4 Gestational Age at Enrollment

Table 2-5 above presents the distribution of the gestational ages at which women claim to have first sought prenatal care. However, that distribution may not be applicable for women going to the centers for such care. Table 2-6 shows the gestational age at which women were enrolled in the Collaborative Prenatal Project (CPP). It can be seen that only about 15 percent of women were enrolled in the first trimester. Since Table 2-5 shows that about 83 percent of women seek prenatal care during the first trimester, it suggests that the collaborating hospitals in the CPP may not be representative of all sources of prenatal care. However, the data from the CPP relate to the period 1959-1965, and much has changed since then. It should now be possible to recruit women earlier in their pregnancies than was the case in the CPP. Exactly how early the enrollments can occur will depend on how the Center draws its participants. For those centers that draw their participants from places of routine prenatal care such as public health clinics, it maybe possible to recruit as early as in the Office Model. For those centers that rely on networks of providers for referrals of pregnant women, recruitment in the Center Model might be somewhat slower than in the Office Model. The mix of patients will also influence the comparison if women who visit the centers for their prenatal care have a different timing for first prenatal care visit than do women across all offices.

Table 2-6. Weeks of gestation at time of registration from the Collaborative Perinatal Project (1/1/1959-12/31/1965)*

Weeks of gestation	Percent	Weeks of gestation	Percent
1-4	0.1	1-12	14.7
5-8	3.3	13-24	47.5
9-12	11.3	25-36	32.7
13-16	14.9	37-40+	5.0
17-20	16.3	Unknown	0.1
21-24	16.3		
25-28	13.3		
29-32	11.0		
33-36	8.4		
37-39	3.8		
40+	1.2		
Unknown	0.1		

*There were 39,215 pregnant women with stillbirths or live births. The table excludes pregnancies that ended prior to the 20th week of gestation. It also excludes multiple births.

2.4.5 Center Timeline

Figure 2-3 shows the projected feasible time line with the Center Model, assuming a project kickoff at the first of 2004, following preparatory research in 2003. With this approach, the total required time is five years—slightly longer than with the Office Model but much shorter than with the Household Model.

2.5 Collection of Pre-Pregnancy Data

The U.S. Government Agencies with principal responsibility for planning the survey have indicated an interest in obtaining pre-pregnancy health and environmental conditions for the women whose children comprise the sample that will be observed over the 20-year followup period. We have not included a discussion of these measurements in our report but have several comments on the feasibility of such an effort.

Three approaches might be employed to integrate pre-pregnancy measurements into the Office or Center Models, but we see very serious problems with each. The first approach we identified would be to obtain information retrospectively for known health conditions prior to pregnancy or known environmental factors. It would probably also be feasible to get signed releases for most women authorizing their medical providers to provide records of past doctor or clinic visits. Some women would undoubtedly refuse, and there may be problems in getting some doctors to cooperate and also in inconsistency in the kinds of records kept, but the overall response should be tolerable. However the analyses of such data could be subject to unknown, and possibly serious, bias arising from the lack of information about medical and environmental conditions the women were not aware of, and which were not reported to their physicians. The second approach would be to recruit women at offices or centers who are seeking routine gynecologic health care or infertility services or even to recruit through broader advertising to women trying to become pregnant. Obviously, such recruitment procedures would skew the sample sharply to older women, better educated women, and women who are generally more concerned about their reproductive health. Such an approach would be very costly and give such a skewed sample that we would not think it worth considering. The third approach would be to sample from some list of women who are likely to become pregnant in future such as those who have had births in the prior 2 years and/or those who have been recently married. These samples would also be substantially skewed (e.g., no first births on the first list, no unmarried mothers on the second) and thus do not seem to worth serious consideration.

The Household Model would make it possible to get pre-pregnancy information. With the Household Model, most of the pregnancies would not be identified at the first visit to the household, but in subsequent contacts over the next three years. It is possible to carry out interviews with all women in childbearing ages at the outset, and take environmental measurements and possibly perform a limited number of health-related tests. There will be approximately 500,000 such women in the screened sample; interviewers must work between two and a half and three hours on the recruitment of each of these women anyway. Adding a 45-minute interview would probably only add $20-$30 million to the project.

Of course, 80 percent will not become pregnant in the time period covered, and the data will not be used for clues to children's health. Health and environmental measurements would cost more.

We have not included this option in our analysis of the features of the various sample designs, or in our cost estimates. The kinds of information that would be desirable, and the measurement techniques, are not yet known. Also, if it were included in the Household Model but not in the other two, it would cause a greater disparity in costs which would not be due to the features of the four designs, but to the fact that the Household Model would be asked to include information not available in the other two models.

2.6 Oversampling Rare Demographic Domains and High-Risk Populations

Although interest has been expressed in the possibility of sampling certain population groups at higher than average rates in order to obtain adequate sample sizes for them, no decisions have yet been made on which, if any, groups should be oversampled. Examination of oversampling strategies was therefore placed outside the scope of our evaluation. Nevertheless, we make some brief observations on this issue.

Oversampling parts of the country such as rural areas or inner cities can be easily accomplished in the area probability sample designs used with the Household and Office Models. All that is required is to oversample PSUs that are located in the parts of the country involved. Oversampling specific demographic domains (such as race/ethnicity) can be partially addressed by oversampling areas where those domains are concentrated, but the effectiveness of such oversampling is lower than is commonly believed (Waksberg, Judkins, and Massey 1997). There would still be a need to increase the overall sample size to generate sufficient sample sizes in the small domains, and screening out superfluous cases in other domains.

Oversampling certain domains in the Center Model can be achieved by setting the desired quota sizes for domains in the selection process. However, the pool of women from which the selections are made would need to be large enough to generate the required domain sample sizes in the given time period. If not, either the Center would need to increase the pool (e.g., by adding extra hospitals to its recruitment network) or by extending the time period.

Interest has been expressed in oversampling of very rare high-risk populations such as fresh-water fish eaters, home gardeners who use pesticides, and workers in certain industries. This type of oversampling would be very difficult and expensive in any of the models since it would require ascertaining the behaviors of a very large sample of women and possibly even analyzing biologics from them to find the women with high levels of the exposures of interest. In the Household Model, the screening sample would need to be made much larger. In the Center Model, networks would have to be expanded greatly or targeted advertising purchased. If information exists about areas with high concentrations of women with the exposures of interest, this information can be used to improve the efficiency of the oversampling, but again following the lines of the oversampling of demographic domains, we would not expect major efficiencies from this approach.

Another consideration in the use of oversampling in a panel survey is that many characteristics change over time. For instance, some of those living in rural areas at the time the sample is selected will move to urban areas during the course of the study and others will move from urban areas to rural areas. Oversampling according to the characteristic at baseline may therefore lose much of its effectiveness later on (unless the baseline characteristic is the analytic measure of interest).

Finally, it should be noted that oversampling certain groups improves the precision of estimates for those groups and of measures of the effect of group membership on, say, health outcomes but these benefits come at the cost of reduced precision of other estimates and measures of effects. A careful evaluation of the various consequences of oversampling is therefore called for.

2.7 Comparisons of the Models

This section briefly compares and contrasts some key features of the three models with respect to the enrollment of pregnant women into the NCS and the conduct of the baseline data collection:

- The Household and Office Models employ probability sampling whereas the Center Model does not.
- For operational reasons, the Center Model has to be restricted to a relatively small number of centers, whereas the samples for the other two models can be more widely spread, providing much broader representation of geographic variety.
- The Household Model requires a large-scale screening operation to identify pregnant women, whereas the other two models do not.
- The Household Model includes a sample of women who do not seek prenatal care, whereas the other two models do not.

- The Household Model can enroll women at an earlier stage of pregnancy than the other two models. The Household Model can even be used to obtain information about women prior to pregnancy, although at substantial additional cost. Pre-pregnancy enrollment with the other models would entail acceptance of highly skewed samples.

- Taking into account both underreporting and nonresponse, the proportion of live births represented in the sample from the Household Model is likely to be in the range of 60 to 65 percent. The corresponding proportion for the Office Model is likely to be smaller, probably in the range of 20 to 30 percent. The concept cannot be calculated for the Center Model since the centers are purposively selected and the women are volunteers.

- The recruitment of pregnant women is spread over five years in the Household Model and three years in the Center Model, whereas, if required, it could be conducted more rapidly in the Office Model.

- The fieldwork involved in recruiting the pregnant women can be more tightly controlled with the Household Model (with professional interviewers) and with the Center Model (with its relatively small number of centers) than with the Office Model (with its large number of offices).

- Assuming that some preparatory work is done in 2003, the total elapsed time including time for everything from developing a management plan to preparing a final methods report is 7.5 years with the Household Model, 4.25 years with the Office Model, and 5 years with the Center Model.

- The Household Model affords the best opportunity for studying miscarriages as women are enrolled earlier in their pregnancies.

- The greater experience of center directors with research procedures may lead to better efforts to conform to study protocols than could be expected in the Office Model. On the other hand, institutional coordination issues are likely to be the most challenging with the Center Model. The same experience that makes it possible to get better conformance to study protocols may make it difficult to agree on those study protocols in the first place. Questions of data rights and IRB issues are also likely to be vexing.

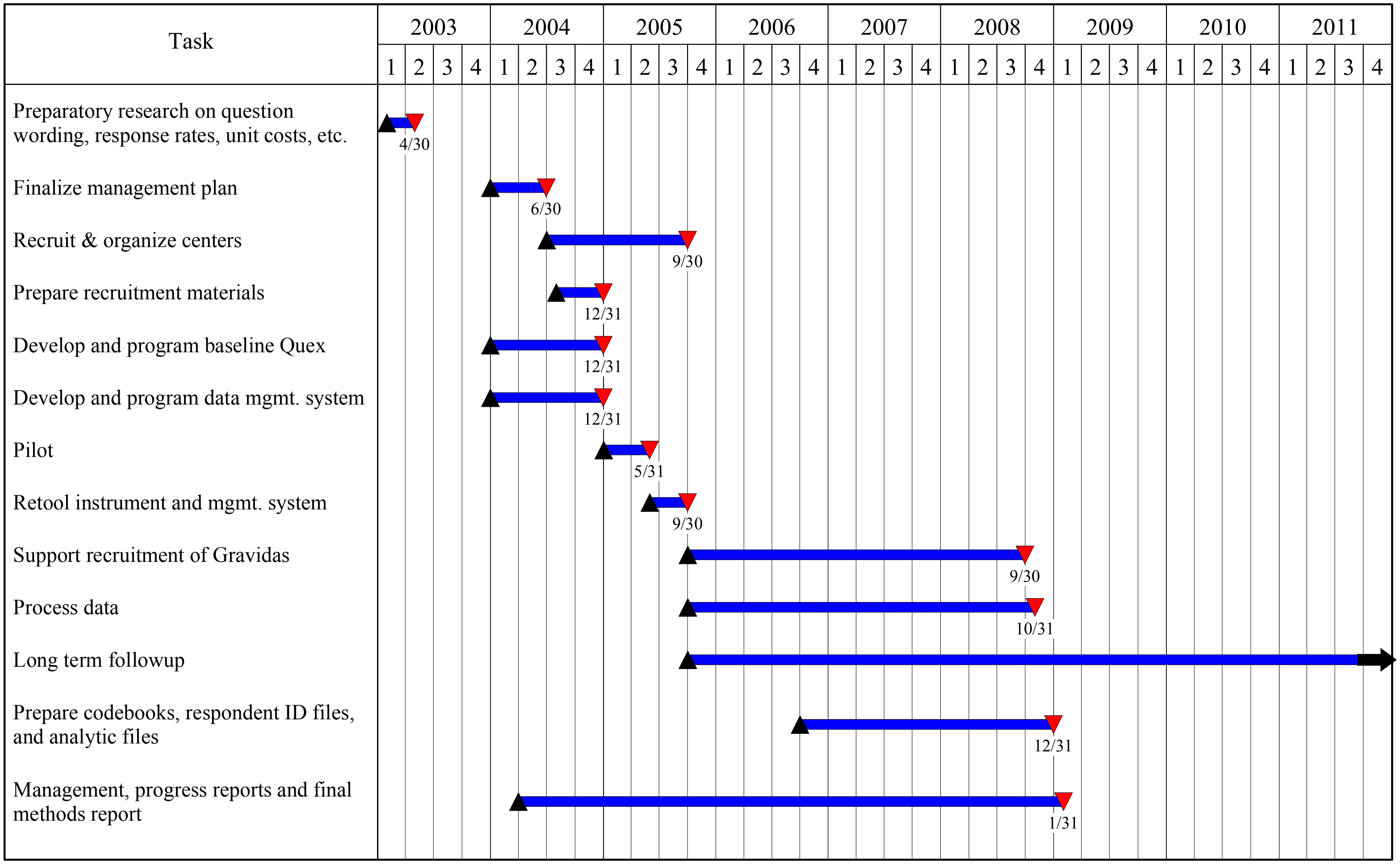

Figure 2-3. Timeline for Center Model

3. ALTERNATIVE SAMPLE DESIGNS

As indicated in Chapter 2, a large number of sample designs are possible with each sampling model. The differences essentially involve the extent of clustering. For both the Household and Office Models, there are two principal levels of clustering: PSUs and segments for the Household Model and PSUs and offices for the Office Model. The extent of clustering depends on how large a sample is taken from each PSU (or, expressed otherwise, how many PSUs are selected) and how the sample of women is distributed geographically within each PSU. (There are also several minor levels of clustering with these two models). For the Center Model, there is only one level of clustering—the centers—and the extent of clustering depends on the number of pregnant women sampled at each center. The ideal practice is to choose what is referred to as an "optimum sample design," i.e., one in which the amount of clustering produces the lowest sampling error for a given estimate for a fixed total cost. This ideal approach, however, cannot be applied exactly to multi-purpose surveys where one narrow statistic (such as the unemployment rate) is not given priority over all other statistics.

There is no single sample design that will provide optimal clustering for a survey with as many objectives as the NCS. The efficiency of a particular clustering plan will depend on a variety of factors, mainly the items of analysis, the importance of information on risk factors that affect all children vs. the effects on subgroups (such as minorities, children in rural areas, those living in areas subject to particular environmental conditions, those growing up in households with various social or economic backgrounds, etc.), the effects of clustering on the cost of followup activities, and the degree to which clusters improves the ability to carry out followup activities for persons who move from their initial sample location.

Nevertheless, the amount of clustering is a major consideration in choosing a sample design. As a practical matter, a design that is optimal for a particular kind of analysis will be efficient for many other analyses. Our choice of designs to consider in detail is therefore heavily dependent on clustering effects that are likely to exist for many of the key variables that are expected to be examined. This chapter discusses the impact of clustering under the three models, and how it affects the choice of sample designs that would be appropriate for the NCS.

3.1 Household Model Sample Designs

The Household Model has four levels of clustering as illustrated in Figure 3-1. One level of clustering is at the PSU level. The second is at the neighborhood level. The third is at the household level, and the fourth is at the mother level. At all levels, clustering tends to reduce cost while increasing variance compared to a simple random sample of the same nominal sample size. An important part of sample design work is to determine the level of clustering that is optimal for a wide range of statistics to come out of the survey.

As indicated in Figure 3-1, there is a choice to be made of what constitutes a PSU. In most national interview surveys, PSUs are defined to be metropolitan areas, groups of counties or large single counties. However, in some surveys, more compact PSUs are needed to facilitate data collection. Thus, for example, the NHANES employs counties as PSUs because it requires each sampled person to go to a mobile examination centers (MEC) for a medical examination; having more compact PSUs reduces the travel time for sampled persons, and this is believed to favorably impact response rates. Since it seems likely that pregnant women and their children in the NCS may also need to travel for some medical examinations under the Household Model, we have opted for the county as the PSU for this model. However, note that the effects of clustering are greater with the more compact definition because of the greater degree of homogeneity of persons in more compact areas (see below).

Given the choice of PSU, the next choice to be made is the nature of the second stage units, termed segments. The usual choice for segments is to create them by combining the blocks identified at the last Census, combining contiguous blocks until they contain sufficient households to yield the desired segment sample size without too great a concentration of sampled households within the segment. In the case of the NCS, however, the population of interest is pregnant women, who comprise only a small fraction of the total population. Thus, concerns about too great a concentration do not arise and it is efficient to screen complete segments for pregnant women. This is the procedure adopted in the Household Model.

As noted earlier, 1.3 million dwelling units need to be screened to produce the required sample of 100,000 live births. With complete screening of sampled segments, the average segment size is then the ratio of 1.3 million dwellings to the number of sampled segments. For example, two of the sample designs for the Household Model considered below contain 12,500 sampled segments. Thus, for these designs, the average segment should be about 105 dwelling units (about five city blocks). Two other Household Model sample designs discussed later are more highly clustered, with only 3,125 sampled segments overall. In these cases the average size of segment needs to be about 416 dwelling units, which is about the size of a Census tract.

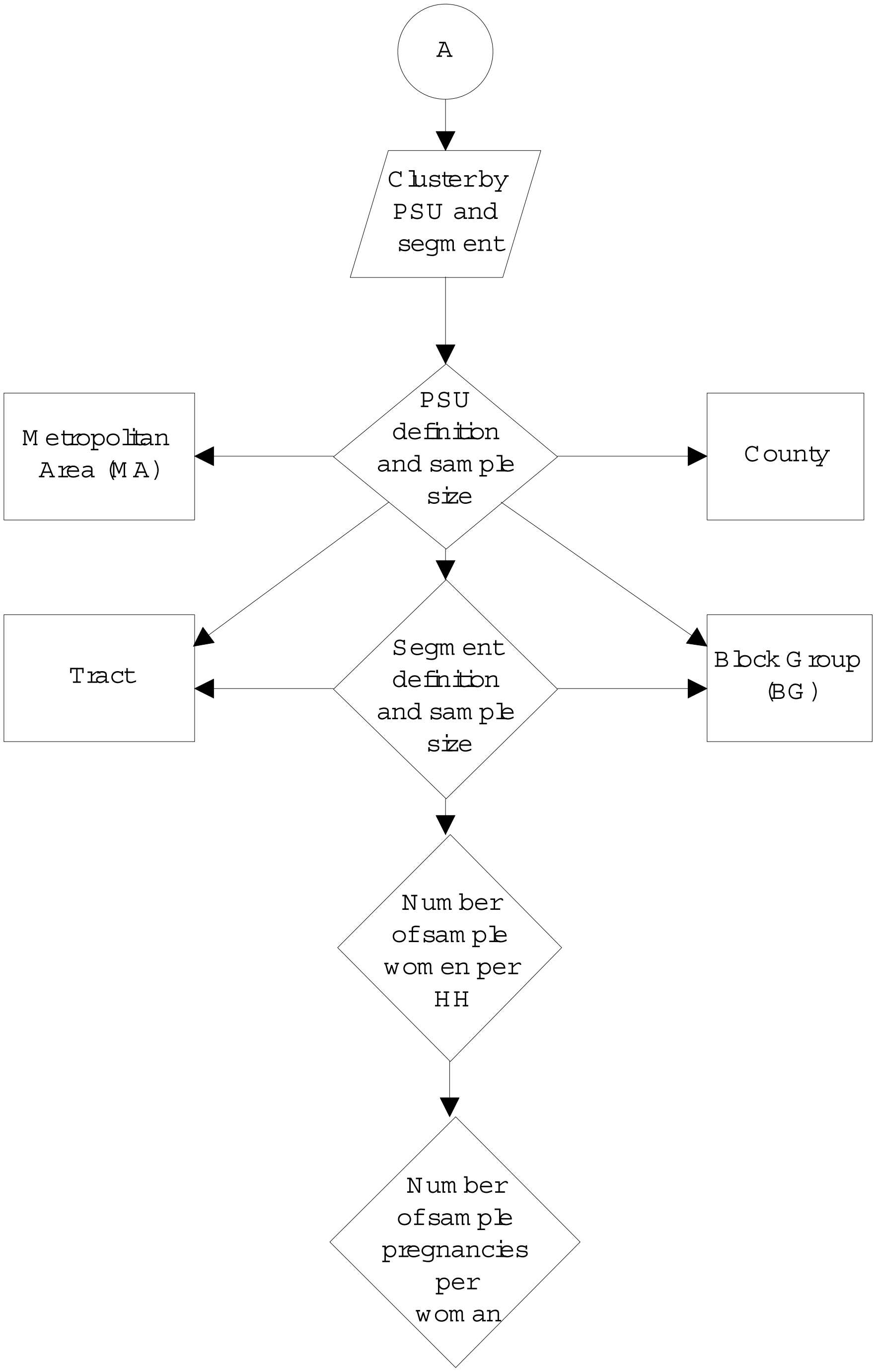
A
Clusterby PSU and segment
PSU definition and sample size
Metropolitan Area (MA)
County
Tract
Segment definition and sample size
Block Group (BG)
Number of sample women per HH
Number of sample pregnancies per woman

Figure 3-1. Clustering Decisions for Household Model Sampling

Following standard practice, the sample design for sampling PSUs is a stratified design in which the PSUs are selected with probability proportional to size (PPS). Stratification is used in order to ensure that the sample of PSUs provides the desired representation of different kinds of PSUs, for example by region, degree of urbanization, minority population, median income. If required, some strata could be sampled at a higher than average rate in order to give adequate sample sizes for persons in the types of areas they contain. The PSUs are sampled by PPS sampling to take account of their unequal population sizes. The use of PPS sampling results in a number of large PSUs being selected with certainty, and particularly so when many PSUs are to be selected, as is likely to be the case for the NCS. Certainty selections will generally be metropolitan counties. The average population size of the 848 metropolitan counties in 2,000 was 267,000 compared with only 24,000 for the 2,293 nonmetropolitan counties. When a PSU is selected with certainty, it is in fact a stratum rather than a PSU. The effects of PSU clustering do not apply to such strata; the only clustering in such strata is at the segment level.

The sample design for segments is also a stratified one to give appropriate representation to different types of segments within each sampled PSU. In this case, however, in view of the plan to take all the pregnant women in a sampled segment into the sample, PPS sampling is not applicable as a means to deal with unequal segment sizes. Instead, the plan is to create segments of as equal a size as possible, and then to take an equal probability sample of them.

The optimal number of sample PSUs will depend on the measurement protocols and the joint distribution of the various exposures and diseases of interest over the geography of the nation. The calculations below are based on some tentative ideas about these factors. However, once analysis issues of particular concern and the measurement protocols have been determined, these calculations should be redone and the number of PSUs should be reexamined.

In a simple case, the effect of clustering at the PSU and segment levels on an estimate of, say, the percentage of children in the total sample with a given condition can be expressed in terms of the design effect (DE) given by

$$\mathrm{DE} = 1 + \delta_{PSU}(\overline{n} - 1) + \delta_{seg}(\overline{\overline{n}} - 1) \tag{3-1}$$

where δ_{PSU} and δ_{seg} are the intraclass correlation coefficients that measure the homogeneity of the condition of interest in the PSUs and segments respectively, and where $\overline{n}$ and $\overline{\overline{n}}$ are the average sample sizes per PSU and per segment respectively.

To illustrate the effect of varying numbers of PSUs and segments on the magnitude of DE, some values are needed for the two intraclass correlations in equation (3-1). The estimates of intraclass correlation for metropolitan areas (MAs) and block groups (BGs) containing about 100 dwelling units given below come from an analysis of Cycles II and IV of the National Survey of Family Growth (NSFG; Waksberg et al., 1993). The estimates of intraclass correlation for counties and other sized segments are extrapolated from these. The MA and BG estimates were generalized from estimated variance components for a wide variety of items studied in the NSFG.

Table 3-1 shows contributions to design effects for different numbers of sample PSUs by whether the PSUs are defined in terms of metropolitan areas (MAs) or counties. A value of $\delta_{PSU} = 0.005$ is assumed when MAs are the PSUs and a value of $\delta_{PSU} = 0.01$ is assumed when the more compact counties are the PSUs. The contribution to design effects from the PSU clustering are computed as the second term on the right of equation (3-1), i.e., $\delta_{PSU}(\overline{n}-1)$. The contributions to DE in these tables assume the items of analysis apply to all children, so that $\overline{n}$ is the average number of children per cluster. For subgroups, $\overline{n}$ will be lower, resulting in smaller design effects.

Table 3-1. Design effect contributions at the PSU level

Number sample PSUs	Babies/PSU	Contribution to DE assuming $\delta_{PSU} = 0.005$ (PSU=MA)	Contribution to DE assuming $\delta_{PSU} = 0.01$ (PSU=county)
100	1,000	5.00	9.99
200	500	2.50	4.99
300	333	1.66	3.32
400	250	1.25	2.49
800	125	0.62	1.24

Table 3-2 shows contributions to design effects for different numbers of sample segments for two levels of intraclass correlation given by $\delta_{seg} = 0.03$ and $\delta_{seg} = 0.015$. These contributions represent the last term in equation (1). To get the total design effect for a design, it is necessary to add the PSU and segment contributions together and then add 1. For example, if there were 800 county-style PSUs and 12,500 segments with $\delta_{seg} = 0.03$, a design effect of 2.49 would result. In simple terms, this means that the sample of 100,000 infants would provide the same variances as a simple random sample of 40,000 infants for analyses relating to the total sample.

Table 3-2. Design effect contributions at segment level

Number of sample segments	Babies/segment	Contribution to DE assuming $\delta_{seg} = 0.03$	Contribution to DE assuming $\delta_{seg} = 0.015$
1,000	100	2.97	1.50
2,000	50	1.47	0.75
3,125	32	0.96	0.48
4,000	25	0.74	0.38
6,250	16	0.48	0.24
12,500	8	0.24	0.12
25,000	4	0.12	0.06
50,000	2	0.06	0.03

As noted earlier, a number of the very large metropolitan counties will be in the sample with certainty. For these PSUs, there is no contribution to the design effect from the PSU clustering. Thus, the PSU clustering will have a greater effect on estimates for populations in nonmetropolitan counties.

It seems likely that a major source of environmental exposures that create health risks in nonmetropolitan counties will be from pesticides in agricultural runoff ingested in drinking water obtained from surface water sources. If there is a large town along a river, it is likely that all the residents in the town obtain their water from the same intake location on the river. As a result, intraclass correlation at the town level at least is likely to be extremely high for water-born environmental contaminants. If there is a series of towns along a river in the same county, the intraclass correlation is again likely to be high even though the intake locations will be different. In fact, it seems likely that within major watersheds, there may be very high intraclass correlations for all the towns along the main river. There may also be a high intraclass correlation for natural contaminants in large aquifers such as radon and arsenic. Based on these considerations, intraclass correlations even larger than 0.5 at the county level do not seem implausible for some exposure variables. The assumption used above of an intraclass correlation of 0.01 at the county level may therefore be a serious underestimate for some items of interest, with the consequence that the design effects will be much larger than indicated for such items. However, since this is all speculative and it is known that disease intraclass correlations are usually much lower, the analysis was done with the 0.01 assumption.

As another source of information on possible levels of intraclass correlation at the segment level, a report on the National Health Interview Survey was consulted (NCHS, 1973). The largest segment size studied in that report was 18 households, roughly one block. The segments for the NCS would be larger and thus have smaller intraclass correlations, but the numbers still give some rough guidance. At one extreme, interclass correlations for being African American (Negro in the 1973 report) was 0.472, a very high level reflecting the residential segregation existing in American society. Air, water, and radiation measurements may have correlations as high or higher than this. At the other extreme, such chronic diseases as tuberculosis and heart diseases actually had negative estimates of intraclass correlations, indicating that spatial proximity was not important. Unemployment and activity limitations had intraclass correlations at the segment level in the range of 0.03 to 0.04. Smoking had an intraclass correlation at the segment level of 0.06.

Another source of estimates of intraclass correlation is the textbook by Hansen, Hurwitz, and Madow (1953, Volume 1, p. 264), which provides a variety of estimates for clusters of size 62 households each. Rent and tenure showed intraclass correlations in the range of 0.06 to 0.11. Unemployment was again between 0.03 and 0.04. Yet another source came from Table 8.1 of a compendium of third world surveys (Le and Verma, 1997). Intraclass correlation at the level of urban neighborhood or rural village ranges from 0.02 to 0.05 for measures of child health and female contraceptive use. Taking these disparate sources into consideration, the assumed segment-level intraclass correlation for NCS of 0.03 seems reasonable.

In addition to the considerable variation that can exist between intraclass correlations at the PSU and segment levels for different variables, and hence different design effects, it also needs to be noted that the design effects reported in the tables above refer to a particular estimate, the proportion of all children with a given characteristic. The effect of clustering is different for different forms of estimates. For example, the design effect for an estimate for a subgroup of children that is spread across the PSUs and segments, such as children in poverty, will be lower than that for the same estimate for all children. Moreover, it may be argued that subgroup design effects are more important since, even with a large design effect, the precision of overall estimates will be high, given the very large overall sample size.

Given that the main aim of the NCS is to discover relationships between various exposures and the development of diseases with long latency periods, it is clear that both the intraclass correlation for exposure and disease will be important. If there is a strong geographic relationship, then the intraclass correlations will tend to be similar, but with weak exposure-disease relationship, the intraclass correlations might be very different.

3.2 Office Model Sample Designs

As shown in Figure 3-2, the levels of clustering consist of PSU, physician and location. With regard to choice of type of PSU, note that with office-based screening, it may be possible to collect the prenatal biologics from the prenatal care offices along with medical records. In this case, the collection of these data would not require the sampled women to travel to other locations, and hence large PSUs such as metropolitan areas would be acceptable. However, these offices are generally not appropriate locations for infant and child examinations, so that the same arguments for more compact PSUs apply as with the Household Model. For this reason, counties are also chosen as the PSUs for the Office Model.[26] The sample of PSUs for the Office Model would thus be selected in the same way as described in Section 3.1 for the Household Model and the contributions of PSU clustering to the design effect with differing numbers of PSUs are the same as those given in Table 3-1.

As described in Section 2-3, the next stages of sampling with the Office Model involve sampling physicians providing prenatal care with addresses within the selected PSUs from AMA lists, sampling the offices of some of the sampled physicians, and sampling pregnant women at the sampled offices. The clustering of the sample within selected offices contributes to the design effect in exactly the same way as does the clustering within selected segments in the Household Model. Equation (1) therefore applies also for the Office Model, with δ_{seg} being replaced by the intraclass correlation within offices $\left(\delta_{off}\right)$ and $\bar{\bar{n}}$ now representing the average sample size per office. We assume here that $\delta_{off} = 0.03$, the same as the higher value assumed for segments. Consequently, the contribution to the design effect from office clustering can be obtained from Table 3-2. However, since we wish to consider a different set of numbers of sampled offices, the design effect contributions from office clustering are displayed in a new table, Table 3-3.

The office design effect contribution from Table 3-3 must be added to that for the PSU contribution taken from Table 3-1 and the quantity 1 must be added to give the overall design effect in equation (1). For example, if there were 4,000 sample offices in 800 county PSUs, the total design effect would be 2.97, meaning that the effective sample size for an estimate based on the total sample would be about 33,700 in contrast to the nominal 100,000.

[26] However, the counties apply to the location of the doctor's office. Some of the patients will live outside the sample counties. They will be included in the sample regardless of where they live.

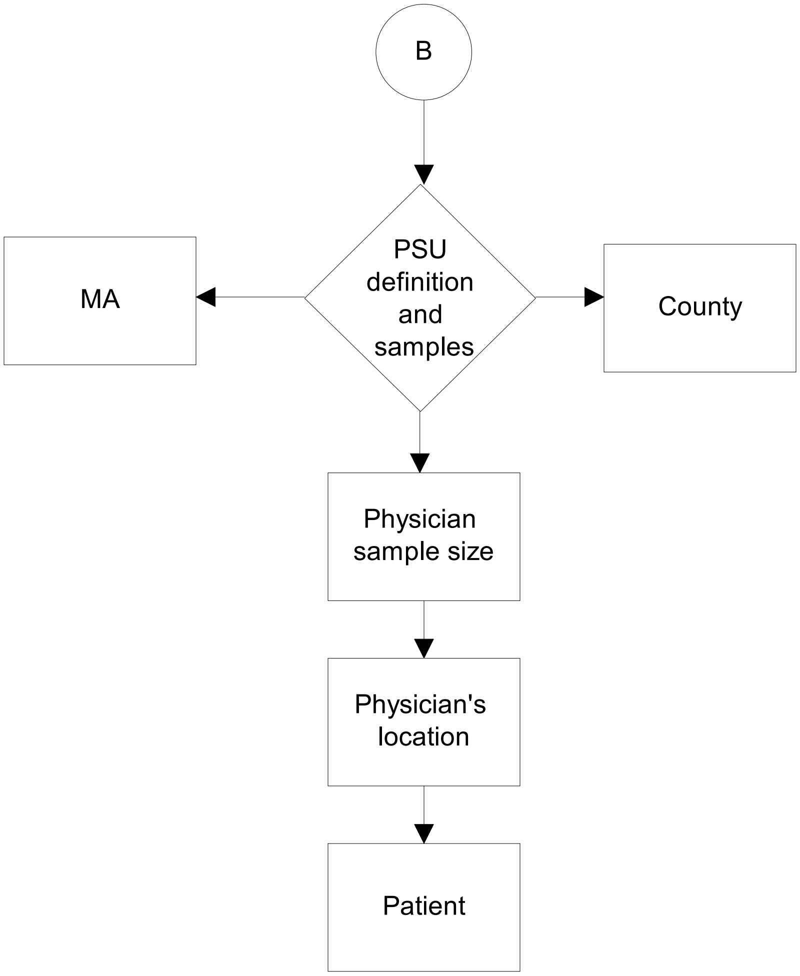

Figure 3-2. Clustering Decisions for Office Model Sampling

Table 3-3. Design effect contribution from sampling offices

Number of sample offices	Babies/office	Contribution to DE assuming $\delta_{off} = 0.03$
500	200	5.97
1,000	100	2.97
2,000	50	1.47
4,000	25	0.72

As discussed for the Household Model, the effect of the clustering will depend on the characteristics being analyzed and the statistic being computed. For environmental exposures about all groups and for most statistics about isolated subgroups such as migrant workers, we would expect more severe variance consequences; while, for many other chronic conditions, it would be reasonable to expect less severe variance consequences for most groups.

3.3 Center Model Sample Designs

Under the Center Model, a set of centers is selected to recruit the sample of pregnant women and to be responsible for collecting all the NCS measurements on them and their children for the duration of the study. The sampling decisions with this model are the selection of centers, and then the processes that a center uses to recruit its sample of pregnant women.

The number of centers selected will necessarily be small. Given the nature of the selection process, the sample is not amenable to the application of survey sampling theory, particularly with regard to the assessment of the bias in the estimates. It is, however, reasonable to examine the consequences of the clustering of the sample in centers using the design effect approach used for the other two models.

For application of the design effect model, we assume only one level of clustering, that is, clustering in centers. The centers can be treated as PSUs and their contribution to the design effect is calculated accordingly. Table 3-4 gives projections of these contributions for varying numbers of centers ranging[27] from 60 to 100 with an assumed intraclass correlation of $\delta_{cen} = 0.01$. As in the other tables of such contributions, the number 1 needs to be added to the contribution in order to get a design effect.

[27] We were advised by some members of the expert panel assembled to review a draft of this report that only about 20 to 30 institutions would be able to perform the role of center. If so, it might be necessary to refashion the Center Model to consist of 20 or 30 center consortia where each

Table 3-4. Design effect contributions from centers

Number of sample centers	Babies/center	Contribution to DE assuming $\delta_{cen} = 0.01$
60	1,667	16.66
80	1,250	12.49
100	1,000	9.99

The design effect contributions in Table 3-4 are based on a projected intraclass correlation of 0.01. This value may well be a serious underestimate. Nevertheless, the design effect contributions are very large, even for a selection of 100 centers. For an estimate of a proportion from the total sample from 100 centers, for example, the effective sample size corresponding to 100,000 children is only about 9,100.

Some of the design effect with the Center Model will come from correlated measurement errors. This model assumes that center staff would not only do the recruiting but would also carry out the measurements after the children are born and as they grow. Depending on the success in getting the centers to standardize their measurement procedures, the design effect contributions from the use of differing procedures may be substantial.

This latter point draws attention to the need for careful training of the data collectors with the Center Model. Given the small number of centers involved, it may be possible to ensure the correct application of highly standardized measurement procedures. However, the disadvantage of a small number of data collectors is that small differences in measurement techniques among them are intensely magnified by their large workloads. Mathematically, the variation in measurements among centers can be considered as contributing an additional component of variance. The design effect resulting from the variation in measurements among a set of centers given by

$$\mathrm{DE} = 1 + (\bar{n} - 1)\frac{\sigma_B^2 - \dfrac{\sigma_W^2}{\bar{n} - 1}}{\sigma_B^2 + \sigma_W^2},$$

where σ_B^2 is the mean square deviation among the center's average subject-level measurement errors, σ_W^2 is the average across centers of each center's individual mean square deviation in measurement

consortium comprises a lead medical center and an affiliated network of local community hospitals. Each consortium would have a single medical director who would leverage personal and professional relationships to ensure that the affiliated local hospitals would follow study protocols.

errors across subjects, and $\bar{n}$ is the average number of subjects per center. Thus although the use of a small number of centers might reduce one or both of the two components of measurement error, these reductions are unlikely to offset the penalty with increasing center sample sizes.

When modeling the relationships between two variables, the calculus is somewhat different from that given above. For this type of analysis, measurement variance in the independent (exogenous) variables leads to bias in estimates of effects, associations, and covariances. Here it is important that the measurement errors be small in magnitude compared to the variation in the outcome associated with the independent variable. In this context there is particular emphasis on σ_W^2, but σ_B^2 can be important also.[28] For these, it is clear that reducing the number of centers might lead to reductions in at least σ_B^2 that would be useful. It is more difficult to see that limiting the number of centers will be useful for reducing σ_W^2. For it, the basic measurement process and the training regimen that are developed are more important.

An additional factor to consider with the Center Model is that it is important to have adequate degrees of freedom for the analyses (assuming that multi-level analyses are used). There is no absolute standard for the adequacy of the degrees of freedom, but comparing the 97.5th percentiles (Table 3-5) of the normal and *t*-distributions with various numbers of degrees of freedom (*df*) is a useful guide. The table starts with 12 *df* since that was the number of centers in the CPP. As one can see, if multi-level modeling had been used, the criterion for assessing a significant result (even after accounting for the much higher estimated standard errors) would have been much tougher to meet. A critical value of $t = 2.18$ for a *P* value of 0.05 is needed rather than the 1.96 that applies when the variance estimate is measured precisely and the Normal distribution is employed.

Summing up these various considerations about sample size, it was felt that appreciably fewer than 100 centers would result in unacceptable variance penalties and unacceptable degrees of freedom. On the other hand, many more than 100 centers might make it impossible to control the variation in measurement levels to be smaller than that the desired effect sizes and thereby lead to unacceptably biased effect estimates. For these reasons, the only Center Model considered below is one with 100 centers.

[28] The between center component might be an important component of the bias in effect estimates if most of the true variation in the independent variable happens to be across centers. For example, suppose one is interested in the relationship of exposure to substance Z to the subsequent development of Condition Y. If exposure to Substance Z occurs in only a few of the centers, then substantial between-center variance in the measurement of Y could lead to substantial bias in the estimated association between Substance Z and Condition Y.

Table 3-5. Critical value for hypothesis tests given various degrees of freedom

Distribution and degrees of freedom (*df*)	97.5th percentile
Student's *t* with 12 *df*	2.18
Student's *t* with 24 *df*	2.06
Student's *t* with 50 *df*	2.01
Student's *t* with 100 *df*	1.98
Student's *t* with 200 *df*	1.97
Normal	1.96

3.4 Sample Designs Considered

As indicated earlier, a large number of specific sample designs, that differ in the amount of clustering, can be considered for each model. At the request of the Government, a set of eight designs was initially considered. Table 3-6 summarizes the features of these eight initially considered designs. As noted in the previous section, only one Center Model design—with 100 centers—was considered. Four Household Model and three Office Model designs were considered, with varying degrees of clustering.

Table 3-6. Possible designs for the National Children's Study

Design	Model	Type of PSU	Type of segment	Number of sample PSUs	Number of sample segments	Screener sample size per PSU	Screener sample size per segment	DE
1	Household	BG†	n/a	12,500	n/a	105	n/a	1.2
2	Household	County	BG†	800	12,500	1,646	105	2.5
3	Household	County	Tract†	100	3,125	13,170	421	11.5
4	Household	County	Tract†	300*	3,125	4,389	421	4.8
5	Office	Office	n/a	4,000	n/a	n/a	n/a	1.8
6	Office	County	Office	800	4,000	n/a	n/a	3.0
7	Office	County	Office	100	500	n/a	n/a	17.0
8	Center	n/a	n/a	100	n/a	n/a	n/a	n/a

*The rural areas will be oversampled with twice as many selected as would normally be chosen. However, the workload within these rural PSUs would be cut in half, so that the overall probability of selection is the same in both urban and rural areas.

†With $\delta_{seg} = 0.03$ for BGs and 0.015 for the larger tracts.

Designs 1 and 5 give widely dispersed samples, under the Household and Office Models respectively, by avoiding the PSU clustering. These designs are attractive for examining the effects of environmental factors that are highly clustered geographically. However, they were rejected from further consideration because of the very high cost and logistical problems associated with data collection for them and because of the belief that most important environmental factors will be only moderately clustered geographically.

Designs 3 and 7 produced highly clustered samples in 100 PSUs for the Household and Office Models respectively. They were both rejected because of the very large design effects projected for them[29]. Some consideration was also given to a variant of Design 3 with 100 PSUs and 12,500 segments, but the design effect would have been reduced from 11.5 down to only 11.1 by such a variant. The general conclusion is that 100 PSUs is not an efficient use of resources with a sample of 100,000 infants.

The remaining four designs were Designs 2, 4, 6, and 8. These designs were selected by the Government for more detailed analysis. They were selected because they illustrate principal features of the three models, not because they are necessarily the "best" design(s) for each model (although they are plausible). A decision on the actual sample design to be used in the NCS and its allocation across sampling stages should not be made until the measurement protocols for health outcomes and environmental exposures as well as priorities on items of analysis have been established.

Features of the four selected designs are given below. Chapters 4, 5 and 6 present power calculations, bias/operational considerations, and cost projections for these designs.

1. **Household Model A**: 800 counties and 12,500 segments. This design requires an initial sample of 1,159,000 occupied residential addresses of which 1,101,000 respond, 622,000 age eligible women, 118,000 pregnancies reported over a three year period, and 100,000 cooperating mothers.

2. **Household Model B**: 300 counties and 3,125 segments. The screening sample sizes for Household Model A apply here also. With this design it is suggested that the rural PSU's would be oversampled to get about twice the number that would be selected without oversampling. However, the workloads within each rural PSU would be cut in half, so that rural children have the same overall probability of selection as other children.

3. **Office Model**: 800 PSUs **and** 4,000 medical provider offices. About five offices would be selected per PSU in most of the U.S., with two to three in the mostly rural counties. Pregnant women would be recruited at their initial visits.

[29] However, the complexities of collecting placentas and cord blood samples, if they are required, might argue in favor of a smaller number of counties with the Household Model.

4. **Center Model**: 100 centers. Each center would recruit 1,000 pregnant women, using a variety of recruitment methods.

4. POWER PROJECTIONS

A common form of analysis of NCS data will be to examine whether a higher level of exposure to a given risk factor is associated with a higher prevalence of a certain condition or disease. In the simplest case, the analysis may involve simply the comparison of the proportions of children with the condition in groups of NCS children exposed and not exposed to the risk factor. This chapter presents some illustrative results of the power of a significance test to detect a difference between two such proportions for a specified level of true difference for the four alternative designs described in previous chapters. A key point to note is that the power depends on the complex sample design employed, and as a rule it is lower than would apply if the NCS children were to be selected by simple random sampling.

4.1 Analysis Methods

A variety of analytic methods—such as regression models, proportional hazard models, and survival models—will be used to analyze NCS data. For purposes of power projection, we focus on the simple comparison of exposed and unexposed subjects, where both sets have been reweighted to balance confounding factors and factors related to nonresponse. Some analytic methods will have somewhat better or worse power, but these simple calculations should still provide a sound basis for planning.

An important feature of any analysis of NCS data is that it should take the clustering of the sample into account. With complex probability sample designs, such as those used in the Household and Office Models, this can be done in a standard way using one of the various software packages for survey data that take account of the sample design, and in particular the clustering involved. The precision of the survey estimates is reduced by the clustering, as is reflected in the design effect discussed in the previous chapter for descriptive estimates. Analytic statistics like regression coefficients and relative risks are also affected by the clustering, as indicated by the treatment in this chapter.

The precision of estimates and the power of significance tests with data from the Center Model are also affected by clustering, and that needs to be reflected in the analytic methods employed. Often the approach used in this kind of case is to employ some form of random effects model—hierarchical or multilevel model—to take account of the clusters (Bryk and Raudenbush, 1992; Longford, 1993; Goldstein, 1995; Hox, 2002). The application of such models is now straightforward given the availability of several packages for computing them. It is important to use such methods in order to avoid the overstatement of the precision of estimates and the understatement of significance levels in significance tests that result from the use of standard methods that fail to reflect the clustering. It is of interest to note here that the same issue of clustering arises in multi-center clinical trials; the need to take the clustering into account in the analysis was noted many years ago by Cornfield (1978) and is emphasized in the recent book by Donner and Klar (2000). The effect of taking the clustering into account through multilevel models is similar to that obtained by applying the survey sampling approach. It results in a design effect like that already discussed. We have therefore applied a design effect with the Center Model in the same way as with the other models.

4.2 Theory for Power Calculations for Comparisons of Exposed and Unexposed Subjects

In general, a large-sample significance test of the difference between the proportions in two disjoint groups is based on a test statistic

$$z = \frac{p_e - p_u}{se\left(p_e - p_u\right)} \tag{4-1}$$

where p_e and p_u are the estimates of the proportions in the two groups and $se\left(p_e - p_u\right)$ is the estimated standard error of the difference in those proportions assuming that the population proportions in the two groups are the same. We will refer to p_e and p_u as domain means. The subscripts '*e*' and '*u*' are used here to denote 'exposed' and 'unexposed' in line with the application under consideration. Provided that the number of sample PSUs is large enough, *z* is approximately normally distributed. With a 5 percent significance level and a two-sided significance test, the difference between the domain means is said to be statistically significant if $|z| > 1.96$.

The above test applies with any sample design. What changes with different designs is the formula for the standard error in the denominator of *z*. In general, the variance of the difference between the domain means $\left(p_e - p_u\right)$—i.e., the square of the standard error—is given by

$$V\left(p_e - p_u\right) = V\left(p_e\right) + V\left(p_u\right) - 2Cov\left(p_e, p_u\right) \tag{4-2}$$

The values of $V\left(p_e\right)$, $V\left(p_u\right)$ and the covariance $Cov\left(p_e, p_u\right)$ depend on the sample design employed.

It is useful to consider first the simple case in which the two samples are selected independently by simple random sampling (SRS) since this case serves as a benchmark for more complex cases. In this case, $Cov\left(p_e, p_u\right) = 0$ since the exposed and unexposed samples are disjoint. Under the null hypothesis,

$$V\left(p_e - p_u\right) = \frac{\pi(1-\pi)}{n_e} + \frac{\pi(1-\pi)}{n_u} \tag{4-3}$$

where π is the proportion with the condition in both groups. Under the alternative hypothesis,

$$V\left(p_e - p_u\right) = \frac{\pi_e(1-\pi_e)}{n_e} + \frac{\pi_u(1-\pi_u)}{n_u} \tag{4-4}$$

where π_e and π_u are the proportions with the condition in the exposed and unexposed groups, respectively. Using these results, the power of the significance test to detect a true difference of $\left(\pi_e - \pi_u\right)$ can be computed for given values of π_e, π_u, n_e and n_u.

With a complex sample design, the two variances in equation (4-2) can be expressed as

$$V\left(p_i\right) = \mathrm{DE}\left(p_i\right) V_o\left(p_i\right) \text{ for } i = e \text{ or } u$$

where $\mathrm{DE}\left(p_i\right)$ is the design effect for p_i and $V_o\left(p_i\right)$ is the SRS variance of p_i under the null or alternative hypothesis as given in (4-3) or (4-4). The design effect arises principally from sampling clusters where people in the same cluster tend to be more similar to each other than would be expected by chance. The reasons for the similarity are usually difficult to disentangle, reflecting as they do the cumulative effects of the complex social interactions of group members. Consider, for example, $V\left(p_e\right)$ under the alternative hypothesis,

$$V\left(p_e\right) = \mathrm{DE}\left(p_e\right) \frac{\pi_e(1-\pi_e)}{n_e} = \frac{\pi_e(1-\pi_e)}{n_e^*}$$

where $n_e^* = n_e / \mathrm{DE}\left(p_e\right)$ is the *effective sample size* for the exposed group. Thus the variance terms in the power calculations can be computed as if simple random sampling had been used, but with effective sample sizes replacing actual sample sizes.

With clustered sample designs, the covariance term in (4-2) is often positive. (See for example, Kalton and Blunden, 1973; Kish and Frankel, 1974.) This positive covariance in domain means arises from the common influence of local factors on both exposed and unexposed subjects when both groups are spread across the sampled PSUs and segments. This covariance will vary by disease and type of exposure. It is difficult to predict in advance of conducting a study, as are the design effects for p_e and p_u. As a result, we have prepared two sets of power calculations where one is more optimistic than the other and we hope to have bracketed the actual power that will be obtained.

Before presenting the results of the power calculations in the next section we set out the assumptions involved in determining effective sample sizes and in computing power. Sets of parameter values are assumed for the illustrative power calculations. It should be noted that these parameter values may be applicable for one condition or disease of interest but not for another. For this reason, the findings cannot be generalized indiscriminately.

4.3 Assumptions for Power Projections

Several assumptions need to be made in order to perform the power calculations. The assumptions made for the power calculations that follow are listed here. The assumptions are important because the results and power are sensitive to them.

- We assume a two-sided significance test with a five percent significance level. If a particular test specified a directional alternative hypothesis and a one-sided test was applied, the power would be considerably improved.
- An overall sample size of 75,000 is assumed, as might be the sample size remaining in the NCS when the children have become teenagers.[30]
- The exposed group is assumed to comprise 20 percent of the sample (i.e., 15,000 youth) and the unexposed group the rest (i.e., 60,000 youth).
- Effective sample sizes are computed to take account of the complex sample design, weighting adjustments to compensate for nonresponse, and adjustments for confounders. The details of these computations are given below.
- A lower bound on power was calculated assuming that exposure is completely segregated by cluster. The rationale for this calculation is discussed below.

[30] This projection is based on four cohort studies from New Zealand, Great Britain and Rhode Island that followed children for many years. These studies had response rates that varied widely from year to year depending upon the methods employed for each followup. We examined these is some detail to develop this projection. The references are Woodward and Fergusson (2001) on the Christchurch Health and Development Study; Wadsworth, Mann, and Rodgers (1992) for the 1946 British Birth Cohort; Power, et al. (1997) for the 1958 British Birth Cohort; and Zornberg et al. (2000) for a subsample of the Rhode Island component of the Collaborative Perinatal Project. Of course, the NCS response rates will depend strongly on: respondent burden; the nature and frequency of contacts; the content of interviews, and of the health and exposure measures; tracing and tracking methods and ability to follow mobile people; and on the skill of the data collectors.

- A likely upper bound on power was calculated assuming that exposure is perfectly uniform across all clusters and the covariance term in equation (4-2) is zero. The rationale for this calculation is also discussed below.

The effective sample size for group i is $n_i^* = n_i / \mathrm{DE}\left(p_i\right)$, where $\mathrm{DE}\left(p_i\right)$ is the design effect for p_i. This design effect is computed by the following general formula for all four designs:

$$\mathrm{DE}\left(p_i\right) = 1 + \delta_{PSU}\left(\frac{\Sigma_j {n_{ij}}^2}{\Sigma_j n_{ij}} - 1\right) + \delta_{seg}\left(\frac{\Sigma_j \Sigma_k {n_{ijk}}^2}{\Sigma_j \Sigma_k n_{ijk}} - 1\right) + 0.12 + 0.23 \tag{4-5}$$

where δ_{PSU} is the intraclass correlation within PSUs, δ_{seg} is the intraclass correlation within segments (i.e., offices for the Office Design), n_{ij} is the sample size for group i in PSU *j*, n_{ijk} is the sample size of group i in segment *k* of PSU *j*, and the additions of 0.12 and 0.23 are allowances for the effects of nonresponse weighting adjustments and adjustments for confounders, respectively. The nonresponse adjustment factor is based on experience from other surveys and the confounder adjustment is based on experience gained in the analysis of the National Survey of Parents and Youth.

The assumptions made for the intraclass correlations are those discussed in Chapter 3: $\delta_{PSU} = 0.01$ for all three models and $\delta_{seg} = 0.03$ for the Household and Office Models (denoted by δ_{off} in Chapter 3). The Center Model assumes no clustering within centers and hence δ_{seg} does not apply.

The magnitudes of the quantities $\Sigma_j {n_{ij}}^2 / \Sigma_j n_{ij}$ and $\Sigma_j \Sigma_k {n_{ijk}}^2 / \Sigma_j \Sigma_k n_{ijk}$ in equation (4-5) depend on the distributions of the exposed and unexposed groups across the PSUs and segments. At one extreme, the domains may be evenly distributed across both PSUs and segments, in which case they are what are known as crossclasses. With crossclasses, n_{ij} and n_{ijk} are approximately constant across PSUs and segments respectively, and hence these two quantities reduce to $\left(n_i / a\right)$, where a is the number of sampled PSUs, and $\left(n_i / b\right)$, where b is the number of sampled segments. Based on these quantities and the above intraclass correlations, the effective sample sizes for the two crossclass groups and the four designs can be computed. They are reported in Table 4-1.

Table 4-1. Effective sample sizes for an exposed crossclass of 15,000 youth and an unexposed crossclass of 60,000 youth for the four illustrative designs

Design	Exposed group	Unexposed group
Household Model		
A: 800 PSUs, 12,500 segments	9,782	27,223
B: 300 PSUs, 3,125 segments	7,677	15,440
Office Model		
800 PSUs, 4,000 offices	9,317	23,904
Center Model		
100 centers	5,282	8,174

The other extreme for the magnitudes of the quantities $\Sigma_j n_{ij}^2 / \Sigma_j n_{ij}$ and $\Sigma_j \Sigma_k n_{ijk}^2 / \Sigma_j \Sigma_k n_{ijk}$ in equation (4-5) occurs when one set of PSUs contains only members of one domain and the remaining PSUs contain only members of the other domain. In this case the domains are termed segregated classes. With segregated classes, under the assumption that the sample size per PSU is constant, the two quantities reduce to n/a and n/b, where n is the total sample size. Using these quantities and the above intraclass correlations, the effective sample sizes for the two segregated classes and the four designs can be computed. They are reported in Table 4-2.

Table 4-2. Effective sample sizes for an exposed segregated class of 15,000 youth and an unexposed segregated class of 60,000 youth for the four illustrative designs

Design	Exposed group	Unexposed group
Household Model		
A: 800 PSUs, 12,500 segments	6,179	24,717
B: 300 PSUs, 3,125 segments	3,311	13,245
Office Model		
800 PSUs, 4,000 offices	5,357	21,429
Center Model		
100 centers	1,697	6,787

Empirical studies such as Kalton and Blunden (1973) have found that the covariance between domain means depends on a number of factors. The covariance is stronger when the domains are more uniformly spread across clusters than when some clusters are much richer in some domains than in others. Also, when the intraclass correlation for the outcome is high, the covariance in domain means tends to be higher. Note that in the case of segregated classes, the exposed and unexposed persons are in different geographic areas so there is no possibility for local factors to influence both. Thus, the covariance between segregated class means is zero.

At this point, we have no good information about how much exposure might vary across the clusters nor how domain means might covary across clusters. We have therefore calculated what we believe to be reasonable upper and lower bounds on the power that will be achieved for various analyses. We have done this by assuming segregated classes as providing a lower bound on power and by assuming crossclasses with zero correlation between crossclass means as providing an upper bound. The lower bound is pretty good since design effects are highest with segregated classes and therefore the effective sample sizes are the smallest, and therefore power is worst.

The upper bound is a bit softer. If exposure was perfectly uniform so as to create crossclasses, and if there was strong intraclass correlation on the outcome measure, then we would expect to see a fairly strong covariance induced between the crossclass means. Such a positive covariance would reduce the variance below the level projected here and hence result in better power than projected. However, as noted above, true crossclasses will not occur in practice, and hence the power results for them are overly optimistic for what can be expected in the NCS. In particular, few if any environmental factors of interest will be evenly spread across both PSUs and segments. Thus ignoring the effect of the covariance term may not be unreasonable as a counterbalance to an uneven distribution of exposure across PSUs.

Note that the computations of effective sample sizes incorporate allowances for nonresponse adjustments and for controlling for confounders in addition to the effects of the complex sample designs used in the four chosen designs.

4.4 Statistical Power Tables

The results of the power calculations based on the formulae and assumptions given above are presented for the four designs in Tables 4-3 to 4-10. Tables 4-3 to 4-6 relate to crossclasses and Tables 4-7 to 4-10 relate to segregated classes. The results in these tables provide general guidance on the likely power that is attained when the two groups are completely evenly spread across the PSUs and segments and when they are completely separated in different PSUs and segments. The first set of tables present rough upper bounds on power while the second set presents rough lower bounds on power. In practice the groups will generally fall between these two extremes, with greater concentrations of one group in some PSUs and segments and of the other group in other PSUs and segments.

In assessing the power results in Tables 4-3 to 4-10, it is worth repeating that the results are based on assumptions about PSU and segment intraclass correlations for one hypothetical condition. These intraclass correlations will vary across conditions, and the variation can be substantial, thus leading to very different design effects and hence very different results for power.

The power of the significance test depends on the proportions with the condition in the two groups, i.e., on π_e and π_u. However, rather than tabulating the power in terms of these two quantities, it is tabulated in Tables 4-3 to 4-10 in terms of overall disease prevalence (i.e., overall proportion with the condition) and relative risk, where the disease prevalence and the relative risk are given by $\left(0.2\pi_e + 0.8\pi_u\right)$ and π_e / π_u, respectively. Power is presented in the tables as a percentage value.

Note that the tables cover only rare diseases, with the highest disease prevalence being 1 percent. Apart from the Center Model and segregated classes with Household Model B (with 300 PSUs), power is high (over 80%) for all designs for a disease prevalence of 1 percent, even with a relative risk of only 1.5, and it will be higher for more common diseases. With a disease prevalence as low as 0.1 percent and a relative risk of 2.5, 80 percent power is attained only for crossclasses in Household Model A. With rarer diseases power will be even lower.

The following observations are based on the results in Tables 4-3 to 4-10:

- Power increases with both disease prevalence and relative risk. With all designs, based on the assumptions made, power exceeds 80 percent for a disease prevalence of 0.55 percent or greater and a relative risk of 2.5 or greater, or for a disease prevalence of 0.3 percent or greater and a relative risk of 3 or greater.

- Power is appreciably greater for crossclasses than for segregated classes. For example, Table 4-3 for Household Model A shows a (conservative) power of 82 percent for a disease prevalence of 0.6 percent and a relative risk of 1.5 in the case of crossclasses. Table 4-7 shows that this power is only 68 percent for the same combination of disease prevalence and relative risk in the case of segregated classes.

- Power increases with a less clustered designs, as can be seen by comparing Table 4-3 with Table 4-4 and Table 4-7 with Table 4-8 for the two versions of the Household Model. For example, the 82 percent power for the disease prevalence of 0.6 percent and the relative risk of 1.5 in Table 4-3 falls to 68 percent in Table 4-4.

- The high degree of clustering in 100 centers with the Center Model has the effect of a considerable loss of power. This effect is particularly strong for segregated classes.

Note that the segregated class model implies here that 20 of the centers contain all the exposed group and the other 80 centers contain all the unexposed group, an unlikely eventuality. (Note also that it is assumed throughout with the Center Model that the clustering will be taken into account in conducting the significance test.)

- A desired property of NCS is that it should be able to detect a relative risk of 1.5 for a disease prevalence of only 2 in 1,000. Based on the assumptions made in the power computations, none of the designs provides adequate power for this purpose. The highest power for the combination of a relative risk of 1.5 and a disease prevalence of 0.2 percent is only a (conservative) 42 percent, which pertains to crossclasses with the less clustered Household Model A; the next highest power is 38 percent, which applies to crossclasses with the Office Model. As described above, the significance test studied incorporates an allowance for reduction in power through the control for confounders. Even if this reduction is removed, the power for the above two designs is only 45 percent. Furthermore, if all the effects of the complex sampling and control for confounders are ignored—so that the design effect is 1 and the two samples are treated as simple random samples using equations (4-3) and (4-4)—the power is only 59 percent. For such a rare disease or condition, only relative risks larger than 1.5 can be detected with high probability with samples of 15,000 exposed and 60,000 unexposed youth.

Table 4-3. Projected power for the Household Model A with 800 PSUs and 12,500 segments: Crossclasses of 15,000 exposed and 60,000 unexposed youth

Disease	Relative risk							
prevalence	1.5	1.6	1.7	1.8	1.9	2	2.5	3
0.05	16	20	24	28	32	36	55	69
0.10	25	31	38	45	51	57	80	91
0.15	33	42	51	59	66	73	92	98
0.20	41	51	61	70	77	83	97	99
0.25	48	60	70	79	85	90	99	100
0.30	55	67	77	85	90	94	100	100
0.35	61	73	83	89	94	97	100	100
0.40	66	78	87	93	96	98	100	100
0.45	71	83	90	95	98	99	100	100
0.50	75	86	93	97	99	99	100	100
0.55	78	89	95	98	99	100	100	100
0.60	82	91	96	99	99	100	100	100
0.65	84	93	97	99	100	100	100	100
0.70	87	95	98	99	100	100	100	100
0.75	89	96	99	100	100	100	100	100
0.80	91	97	99	100	100	100	100	100
0.85	92	97	99	100	100	100	100	100
0.90	94	98	100	100	100	100	100	100
0.95	95	98	100	100	100	100	100	100
1.00	95	99	100	100	100	100	100	100

Table 4-4. Projected power for the Household Model B with 300 PSUs and 3,125 segments: Crossclasses of 15,000 exposed and 60,000 unexposed youth

Disease	Relative risk							
prevalence	1.5	1.6	1.7	1.8	1.9	2	2.5	3
0.05	13	15	18	21	24	27	42	55
0.10	19	24	29	34	39	44	66	80
0.15	25	32	38	45	52	58	81	92
0.20	31	39	48	56	63	70	90	97
0.25	36	46	56	64	72	78	95	99
0.30	42	53	63	72	79	85	97	100
0.35	47	59	69	78	84	89	99	100
0.40	52	64	75	83	89	93	99	100
0.45	56	69	79	87	92	95	100	100
0.50	60	73	83	90	94	97	100	100
0.55	64	77	86	92	96	98	100	100
0.60	68	80	89	94	97	99	100	100
0.65	71	83	91	96	98	99	100	100
0.70	74	86	93	97	99	99	100	100
0.75	77	88	94	97	99	100	100	100
0.80	79	90	95	98	99	100	100	100
0.85	82	91	96	99	100	100	100	100
0.90	84	93	97	99	100	100	100	100
0.95	86	94	98	99	100	100	100	100
1.00	87	95	98	99	100	100	100	100

Table 4-5. Projected power for the Office Model with 800 PSUs and 4,000 offices: Crossclasses of 15,000 exposed and 60,000 unexposed youth

Disease	Relative risk							
prevalence	1.5	1.6	1.7	1.8	1.9	2	2.5	3
0.05	15	19	22	26	30	34	52	66
0.10	23	29	36	42	49	54	77	89
0.15	31	40	48	56	63	70	90	97
0.20	38	49	58	67	74	81	96	99
0.25	45	57	67	76	83	88	98	100
0.30	52	64	74	82	88	92	99	100
0.35	58	70	80	87	92	95	100	100
0.40	63	75	85	91	95	97	100	100
0.45	68	80	88	94	97	98	100	100
0.50	72	84	91	96	98	99	100	100
0.55	76	87	93	97	99	99	100	100
0.60	79	89	95	98	99	100	100	100
0.65	82	91	96	99	99	100	100	100
0.70	85	93	97	99	100	100	100	100
0.75	87	95	98	99	100	100	100	100
0.80	89	96	99	100	100	100	100	100
0.85	90	97	99	100	100	100	100	100
0.90	92	97	99	100	100	100	100	100
0.95	93	98	99	100	100	100	100	100
1.00	94	98	100	100	100	100	100	100

Table 4-6. Projected power for the Center Model with 100 centers: Crossclasses of 15,000 exposed and 60,000 unexposed youth

Disease	Relative risk							
prevalence	1.5	1.6	1.7	1.8	1.9	2	2.5	3
0.05	10	11	13	15	16	18	28	37
0.10	13	16	19	23	26	30	47	61
0.15	17	21	26	31	36	40	62	77
0.20	21	26	32	38	44	50	73	87
0.25	25	31	38	45	52	59	82	93
0.30	28	36	44	52	59	66	88	96
0.35	32	41	50	58	66	72	92	98
0.40	35	45	55	63	71	78	95	99
0.45	39	49	59	68	76	82	97	99
0.50	42	53	64	73	80	86	98	100
0.55	45	57	68	77	84	89	99	100
0.60	48	61	71	80	86	91	99	100
0.65	51	64	75	83	89	93	100	100
0.70	54	67	78	85	91	95	100	100
0.75	57	70	80	88	93	96	100	100
0.80	60	73	83	90	94	97	100	100
0.85	62	75	85	91	95	97	100	100
0.90	65	78	87	93	96	98	100	100
0.95	67	80	88	94	97	99	100	100
1.00	69	82	90	95	98	99	100	100

Table 4-7. Projected power for the Household Model A with 800 PSUs and 12,500 segments: Segregated classes of 15,000 exposed and 60,000 unexposed youth

Disease	Relative risk							
prevalence	1.5	1.6	1.7	1.8	1.9	2	2.5	3
0.05	14	17	20	24	27	30	46	58
0.10	20	25	31	36	41	47	68	81
0.15	26	33	40	47	54	60	82	92
0.20	32	41	49	57	64	71	90	97
0.25	38	47	57	65	73	79	95	99
0.30	43	54	64	72	79	85	97	100
0.35	48	59	70	78	84	89	99	100
0.40	52	65	75	83	88	92	99	100
0.45	57	69	79	86	91	95	100	100
0.50	61	73	83	89	94	96	100	100
0.55	64	77	86	92	95	98	100	100
0.60	68	80	88	94	97	98	100	100
0.65	71	83	91	95	98	99	100	100
0.70	74	85	92	96	98	99	100	100
0.75	77	87	94	97	99	99	100	100
0.80	79	89	95	98	99	100	100	100
0.85	81	91	96	98	99	100	100	100
0.90	83	92	97	99	100	100	100	100
0.95	85	93	97	99	100	100	100	100
1.00	87	94	98	99	100	100	100	100

Table 4-8. Projected power for the Household Model B with 300 PSUs and 3,125 segments: Segregated classes of 15,000 exposed and 60,000 unexposed youth

Disease	Relative risk							
prevalence	1.5	1.6	1.7	1.8	1.9	2	2.5	3
0.05	12	13	15	17	19	22	32	41
0.10	15	18	21	25	28	31	48	61
0.15	18	22	27	31	36	41	60	74
0.20	21	27	32	38	43	49	70	84
0.25	24	31	37	44	50	56	78	90
0.30	28	35	42	50	56	63	84	94
0.35	31	39	47	55	62	68	89	96
0.40	34	43	51	60	67	73	92	98
0.45	37	46	56	64	71	77	94	99
0.50	40	50	59	68	75	81	96	99
0.55	42	53	63	72	79	84	97	100
0.60	45	56	66	75	82	87	98	100
0.65	48	59	70	78	84	89	99	100
0.70	50	62	72	80	87	91	99	100
0.75	53	65	75	83	89	93	99	100
0.80	55	67	77	85	90	94	100	100
0.85	57	70	80	87	92	95	100	100
0.90	59	72	82	88	93	96	100	100
0.95	62	74	83	90	94	97	100	100
1.00	64	76	85	91	95	97	100	100

Table 4-9. Projected power for the Office Model with 800 PSUs and 4,000 offices: Segregated classes of 15,000 exposed and 60,000 unexposed youth

Disease	Relative risk							
prevalence	1.5	1.6	1.7	1.8	1.9	2	2.5	3
0.05	14	16	19	22	25	28	42	54
0.10	19	23	28	33	38	42	63	77
0.15	24	30	37	43	49	55	77	89
0.20	29	37	45	52	59	65	86	95
0.25	34	43	52	60	67	73	92	98
0.30	39	49	58	67	74	80	95	99
0.35	43	54	64	73	80	85	97	100
0.40	47	59	69	77	84	89	99	100
0.45	51	63	74	82	88	92	99	100
0.50	55	68	78	85	90	94	100	100
0.55	59	71	81	88	93	96	100	100
0.60	62	75	84	90	94	97	100	100
0.65	65	78	86	92	96	98	100	100
0.70	68	80	89	94	97	98	100	100
0.75	71	83	90	95	98	99	100	100
0.80	73	85	92	96	98	99	100	100
0.85	76	87	93	97	99	99	100	100
0.90	78	88	94	98	99	100	100	100
0.95	80	90	95	98	99	100	100	100
1.00	82	91	96	98	99	100	100	100

Table 4-10. Projected power for the Center Model with 100 centers: Segregated classes of 15,000 exposed and 60,000 unexposed youth

Disease	Relative risk							
prevalence	1.5	1.6	1.7	1.8	1.9	2	2.5	3
0.05	10	11	12	14	15	16	23	29
0.10	12	13	15	18	20	22	32	42
0.15	13	16	18	21	24	27	41	53
0.20	15	18	21	25	28	32	48	62
0.25	17	20	24	28	33	37	55	69
0.30	18	23	27	32	37	41	61	75
0.35	20	25	30	35	40	46	67	80
0.40	22	27	33	39	44	50	71	84
0.45	23	29	36	42	48	53	76	88
0.50	25	31	38	45	51	57	79	90
0.55	27	34	41	48	54	60	82	92
0.60	28	36	43	51	57	64	85	94
0.65	30	38	46	53	60	67	87	95
0.70	31	40	48	56	63	69	89	96
0.75	33	42	50	58	66	72	91	97
0.80	34	44	52	61	68	74	92	98
0.85	36	45	55	63	70	77	94	98
0.90	37	47	57	65	72	79	95	99
0.95	39	49	59	67	74	80	96	99
1.00	40	51	61	69	76	82	96	99

5. BIAS AND OPERATIONAL CONSIDERATIONS

This chapter considers some broad issues for the NCS. Section 5.1 discusses issues of potential bias in the NCS results arising from deficiencies in sampling methods. The data collected in most surveys of human populations are subject to some bias, and the NCS is no exception. Biases will be present both in estimates of the distribution of exposure across the population and in estimates of the association between exposure and children's health. Given that the analytic objectives of the NCS place greater importance on estimated associations than exposure distributions, biases in association estimates are a more serious concern than those in exposure distributions. Substantive biases from sampling deficiencies are less likely with association estimates than with estimated exposure distributions. Nonetheless, associations can vary across subpopulations. If associations do vary across subpopulations but this variation is poorly understood, it can be useful to estimate a population-average association for such purposes as cost-benefit analysis in environmental regulation. Deficiencies in sampling methods can lead to important biases in the population-average associations.

Section 5.2 begins with a discussion of the problems in the control of survey operations and consistency of measurements of health conditions taken during pregnancy and over the course of the next 20 years. This includes some discussion of biases in associations caused by errors in the person-level measurements of exposure and health. It then discusses issues of data collection for the human and physical environmental measurements, both in the short- and long-term. Although firm plans have not yet been developed for those measurements, it is useful to make a preliminary assessment of the measurement constraints imposed by each sampling model.

5.1 Bias Considerations

Potential biases can arise from undercoverage of pregnant women in the frame used for sample selection and in the survey frame-building operations, and from nonresponse. Taking both these factors into account, the issue is the extent to which the sample does not resemble the entire U.S. population of children. The different models have varying strengths and weaknesses with respect to sources of bias. Note that the biases are associated with the choice of sampling model, and not with the specific sample design chosen for each model.

This section has three subsections on "undercoverage," "nonresponse," and "representativeness." These distinctions are somewhat artificial in that undercoverage and nonresponse both reflect a lack of representativeness that can occur when using probability sampling. Nonetheless, it seems useful to discuss undercoverage for the Home and Office Models separately from the issue of representativeness in the Center Model.

5.1.1 Undercoverage

It is not possible to determine in advance the effect of undercoverage biases on the conclusions drawn from analyses of NCS data. Undercoverage of children will usually affect estimates of the numbers of children at risk from specific environmental factors more than estimates of the magnitudes of the risks. Nevertheless, undercoverage can also affect the survey's estimate of the magnitude of a risk, and even in some cases the assessment of whether a risk exists.

The Office Model fails to cover women who do not seek prenatal care, and the Center Model either fails to cover these women or at least underrepresents them. Thus, with these models, estimates of the numbers at risk could be seriously biased when the risk is also related to race/ethnicity and socio-economic conditions, since minority and low-income women will be overrepresented in those who do not seek prenatal care. Equally, the estimate of the magnitude of a risk could be biased if the environmental factor is related to socio-economic conditions, or there are other similarities among women who choose not to seek prenatal care.

The term coverage is strictly undefined for the Center Model, given the purposive procedure used to select the centers and the ways the centers will recruit the samples of pregnant women. It is impossible to determine whether the women and their children recruited in this manner are likely to respond to the risk factors in different ways than the rest of the U.S. population.

In addition to failing to cover pregnant women who do not seek prenatal care, the Office Model is subject to undercoverage to the extent that women seeking prenatal care go to physicians not on the AMA list for the specialties from which the physicians are sampled. It also is subject to undercoverage if the list of offices for sampled physicians is incomplete and if the full set of pregnant women attending the sampled offices for the first time in the defined period is incomplete. Undercoverage for this last reason is of particular concern because the identification of the women will be the responsibility of office staff, who may lack adequate commitment to the study.

Like most population surveys, the Household Model is restricted to the civilian noninstitutionalized household population (i.e., excluding women in the military, women in prisons, psychiatric hospitals, etc., and homeless women). For that population, the frame is theoretically complete. Unlike the Office and Center Models, it includes all pregnant women, including those who never seek prenatal care. However, in practice there is bound to be some slippage in coverage. Some households will be missed in the household listing operation in the sampled segments, some age-eligible women will be missed in the sampled households, and some women will deny their pregnancies as an easy way of avoiding participation in the survey, or for other reasons. The use of well-trained listers and professional survey interviewers, together with the standard careful control procedures for the survey operations, will keep the first two of these sources of undercoverage to a low level, and the main concern is the underreporting of pregnancies. Although any amount of undercoverage could affect some of the results, the women missed in the Household Model are unlikely to have as much in common as women without prenatal care, so that the effect of Household Model undercoverage on analyses of risk factors will in most cases be lower.

It should be noted that attempts to compensate for undercoverage can be made in the analysis. Poststratification and other forms of calibration adjustment are regularly used for this purpose in survey research, using population data available from other sources. A valuable data source for making these adjustments for the NCS will likely be national birth certificate data. This rich data source should be very helpful, but nevertheless attempts to compensate for undercoverage at the analysis stage should be regarded as only a partial solution to the problem.

5.1.2 Nonresponse

The effects of low response rates on the survey results are similar to the effects of undercoverage. Estimates of totals will usually be subject to greater bias than estimates of risk. Both types of statistics will be seriously affected if some common attribute of mothers or children is a major cause of nonresponse.

To reach the stage of recruiting pregnant women into the NCS with the Household Model, the first stage of potential nonresponse is at the point of initial household nonresponse. Next comes potential nonresponse in recruiting age-eligible women into the panel that is followed to identify pregnancies in the three year period. After that, some women may drop out or be otherwise lost to the panel prior to the attempt to recruit those who become pregnant into the NCS. With the Office Model, the first stage of potential nonresponse is the recruitment of the sampled physicians and their office staff. Past experience with gaining doctors’ cooperation is not encouraging. Whether NCS can find ways to encourage a high proportion of doctors to participate, including finding persuasive incentives, is a serious concern with the Office Model. When a sampled physician refuses to participate, a substitute physician in the same PSU may be selected as a replacement. This form of nonresponse compensation may well be appropriate but, in common with all forms of compensation for nonresponse, it does not eliminate the problem.

Once pregnant women have been sampled, nonresponse will occur at the initial recruitment and then further nonresponse attrition will occur during the long-term followup over the next 20 years. Willingness to participate at the initial recruitment will mostly be due to such factors as the women's sense of civic responsibility, their perception of their own time pressures, and concerns that physical examinations will traumatize the children or endanger their health. Some nonresponse will also occur because of such factors as failing to find the woman at home and failing to locate the woman.

We believe that response rates at the recruitment stage will be higher in the Household Model than the Office Model, although some offsetting factors could favor the Office Model. The interviewers in the Household Model can be trained on ways of converting reluctant respondents and can be subject to close supervision of their activities. It seems unlikely that staffs in medical provider offices will be as impressed by the need to get high response rates, and will be willing to jeopardize their relationships with clients by urging them to do something that they initially refuse to do. However, doctors' offices do add an aura of legitimacy to the request for participation, and in some cases this may encourage women to participate in the NCS. This positive aspect of the Office Model will be counterbalanced by the nonresponse arising from refusal by medical providers to participate in the study, as noted above. Unless doctor cooperation is high, this source of nonresponse is likely to dominate the nonresponse concerns with the Office Model.

As in the case of undercoverage, nonresponse in the Center Model is essentially undefined. Once recruited, the sample of pregnant women in the Center Model will be subject to attrition nonresponse at subsequent waves of data collection, as is the case with the other models also.

Over the course of the initial pregnancy and then the next 20 years, there is bound to be considerable attrition in the sample. Reasons for attrition will include mothers (and later children) who become tired of being contacted and having periodic examinations, women and children who move to other residences and cannot be traced, and those who move to locations where it is difficult to carry out the subsequent data collections, particularly physical examinations and environmental exposure measurements. We assume that analysts will utilize baseline and intermediate information on attritors in their analyses to reduce the risk of nonresponse bias, but residual biases will probably persist.

The organization of followup activities has not yet been crystallized for the Household and Office Models, but there is no reason why they would differ from each other. The structure and plans for followup will influence response rates in the long-term followup more than the frame used. The number of women and children who drop out for personal reasons will probably differ only slightly among the three frames. The Center Model may have a small advantage since those women who volunteered initially presumably had interest in the research while the other two models include women who may have been reluctant but were persuaded to cooperate. Also, in the Center Model, the relationships the women establish with a well-known, and likely prestigious, medical center may serve to foster continuing cooperation.

Movers will be a source of nonresponse in all three models. Sample designs with a large number of PSUs will have an advantage since more of the women and children will arrive at locations close enough to one of the sample PSUs for them to be included in the PSUs physical examinations and environmental program. As an illustration, the 1984 Survey of Income and Program Participation (SIPP) had 174 PSUs. In that panel, 96.7 percent of movers moved to within 100 miles of one of the sampled PSUs, and could therefore be interviewed in person by the SIPP interviewer at their new location. By exchanging respondents between data collection staff in this manner, following movers should not present too serious a problem for the widely dispersed Household and Office Models. This problem is of greater concern for the Center Model.

Women and children who move without a forwarding address present a tracing problem. A number of steps are commonly taken in panel surveys to help with tracking and tracing sample members. These methods include, for example, obtaining names, addresses, and telephone numbers of persons who will know where sampled people have moved to, sending mailings to the respondents between data collections with address changes requested from the Post Office, and a range of tracing methods if tracking methods fail. The use of such methods should be very effective in maintaining contact with sample members for all three models. It is possible that tracing may be somewhat easier to perform and more effective with the Center Model in view of the close relationship between the centers and sample members established under that model.

We should like to note that there is very little guidance in the research literature on when nonresponse is likely to have a major impact on the results. As a precautionary measure, most experienced survey practitioners follow the practice of putting a fair amount of resources into efforts to achieve reasonable response rates and to collect partial information on nonrespondents for use in making nonresponse adjustments to the survey weights to reduce the risk of nonresponse bias.

5.1.3 Representativeness

We use the term “representativeness” to denote the overall ability of a sample to mirror the population.[31] The Household Model is the best choice with respect to the ability to generalize results so they can be presumed to apply to almost all children in the United States, or all in a particular subset that is the subject of analysis. The Office Model starts off as a probability sample, but the likely large number of physicians who will refuse to participate and the appreciable number of pregnant women who do not seek prenatal care will create distortions in the sample in unmeasurable amounts and directions. Nonresponse and undercoverage will create similar problems in the Household Model, but they are expected to be much smaller.

The Center Model starts off as a nonrepresentative collection of women and children, and will remain so. The analyses will rely heavily on models to compensate for the lack of representativeness. If the centers provide free medical care, this could be a powerful incentive to the medically indigent to enroll in the NCS. Financial incentives also work well for low-income persons. These kinds of incentives could tilt the sample to overrepresent children below, or near, the poverty level. It is unlikely that this frame will provide the same mix of women at different economic classes as exists in the population at large, or have the appropriate geographic distribution. Any conclusions drawn are likely to be greatly influenced by the models used in the analyses.

As noted in Section 5.1.1, the biases caused by a lack of representativeness will be more serious for estimating exposure distributions than disease-exposure associations. The Center Model cannot produce valid estimates of exposure distributions that are representative of the U.S. population. Disease-exposure associations would be less affected by a lack of representativeness, but population-average associations could be seriously biased.

A specific problem of representativeness will occur if the effects of environmental exposures during the early stage of pregnancy (e.g., the first trimester) are desired. In this case, the data will rely on women who are recruited early enough in their pregnancy to permit environmental and other relevant measurements to be taken by the 12th week. The Household Model may be able to identify a substantial proportion of pregnant women in the first trimester, but this is less true for the other models (see Sections 2.3.4 and 2.4.3). There would therefore be considerable uncertainty about generalizing

[31] With weighting adjustments for unequal selection probabilities in general, and particularly so if some groups of specific interest are oversampled.

findings based on women who seek early prenatal care in the Office Model and especially in the Center Model to all pregnant women.

5.2 Data Collection Issues

This section has five subsections. The first discusses measurement systems generally. The remaining subsections are each organized around a major measurement subsystem. Subsystems may be required for human measurements on pregnant women, environmental measurements in the homes and neighborhoods of pregnant women, tissue collection (placentas and chord blood) in delivery rooms, and human measurements on the children as they develop. There might also be a need for continued environmental measurements in the homes of the children.

5.2.1 Control of Survey Operations and Consistency of Measurements

The vast bulk of the measurements, both of the child and the physical and family environment, will take place over the 20 years of the long-term followup. Clearly, the centers will carry out most, if not all, of the physical measurements if the Center Model is used. A medical data collection system will need to be established if either the Household or the Office Model is chosen. (Since the Office Model will mainly use medical offices specializing in obstetrics, we assume they cannot be used for the long-term followup.) There seems to be no reason why the procedures established for medical examinations should differ between the Household and Office Model.

When the National Health and Nutrition Examination Survey (NHANES) was faced with the issue of determining the best way of carrying out physical and medical measurements for a national sample of the population, three alternatives were considered:

- To use each sample individual’s personal physician;
- To arrange for a single medical office in each PSU to conduct all of the examinations needed in that PSU; and
- To establish a set of mobile medical examination centers (MECs) and a staff that was trained to operate these centers, and have the MECs and their staffs move from PSU to PSU.

The third alternative was chosen. The main reason for this choice was that, with much greater ability to train staff to take measurements in a uniform way, this system seemed to provide greatest assurance of consistency in measurements in all PSUs.

There are certainly major differences between the NCS and NHANES, but the experience is relevant and should be taken into consideration. It is important to ensure that the medical data—and, indeed, all the survey data—are collected in a uniform way. Standardizing the medical data collection process is likely to be easier with the Center Model, with the relatively small number of centers involved. The ability to achieve standardization with the Household and Office Models will depend on the methods used to collect the medical data under those models. However, with the widely dispersed samples in these models, achieving adequate standardization is likely to prove difficult for at least some measurement procedures.

5.2.2 Considerations for the Measurement and Collection of Biologics from Pregnant Women

In the short term, the emphasis is expected to be on the anthropometric measurement of pregnant women and the collection of maternal biologics as providing the best measurements of fetal environment. In both the Office and Center Models, these data collections are likely to be fairly simple. In the Office Model, it should be possible to get very high rates of consent if the office physician collects the medical data. Since the mother has agreed to be in the program and the physician is already performing a variety of sensitive procedures, resistance to having other tests carried out will probably be low. obviously easiest to do for those pregnant women who are enrolled from those attending the center for prenatal care. In that model, cooperation rates are likely to be lower for those recruited from other sources, since these women would need to travel to a different facility than they use for routine prenatal health care and undergo examinations by less familiar faces. Problems may be slightly greater in the Center Model if the data collections take place in a different medical facility than where the woman receives her normal prenatal care, but it would probably be easy to develop good solutions including the possibility of defining enrollment to encompass providing the biologics.

Under the Household Model, there would be at least four options for anthropometric measurements and collecting biologics. One would be to have phlebotomists visit each sample woman, which is expensive but not impossible. Since there would be on average 148 pregnant women per PSU to be examined over a three year period, there would only be about one blood draw needed per PSU per week. This would require the part-time services of a phlebotomist in each PSU. These services could be provided by hiring a part-time phlebotomist in each PSU, by hiring a full-time phlebotomist who also collects other data (such as anthropometric measurements on the infants, administration of dietary questionnaires, etc.), or by hiring a full-time phlebotomist who travels from PSU to PSU. The phlebotomists could be trained to also take the anthropometric measurements.

In addition to considerations about arranging adequate work for a visiting phlebotomist, there might be some restraints on the tests that could be run on the samples. Any tests that required the rapid centrifugation of blood or the immediate separation of cells and serum would probably not be feasible unless the phlebotomist could bring a small scale laboratory in his or vehicle. Again, this would add considerably to expense.

A second option for collecting anthropometrics and biologics in the Household Model would be to have the pregnant women go to their regular doctors (if they have them) with a referral order that would specify the measurements to be taken and the biologics to be collected, instructions on invoicing, and any special instructions for storage, shipping, preprocessing, etc. For those pregnant women who have no regular doctor, a cooperating doctor would need to be identified.

A third option for collecting biologics in the Household Model would be to request all the pregnant women in a PSU to go to a particular doctor who had been recruited for the study in that PSU. As with phlebotomists, there would not be much work for the doctor in the initial recruiting stage of the project, but he/she could also be responsible for the more intensive followup activities.

Finally, for the anthropometric measurements in the Household Model, it would be possible to collect and abstract medical records from their regular obstetricians, at least for those women who receive prenatal care.

5.2.3 Short-Term Considerations for Physical Environmental Measurements

We assume that the project will require the collection of air, dust and water samples within the homes of pregnant women or in their neighborhoods. Radiation measurement is also a possibility. The quality of neighborhood social functioning has also been mentioned but with no clear measurement protocol. At this time it is not clear what equipment will be needed for the environmental measurements or the level of staff training that will be required to operate the equipment.

If the pregnant women can collect the environmental samples themselves and mail them into a laboratory, then all the sample designs are equal in terms of the cost of the measurements. If, on the other hand, the environmental measurements must be taken by professional environmental engineers or other trained staff, then the different sample designs pose somewhat different considerations. In this case, the Household Model has a slight advantage since the pregnant women will be clustered in sampled segments. However, the spread of the recruitment over three years means that the clustering may not be very helpful for measurements of specific households. Still, the trained staff could conduct neighborhood measurements at the start of the recruitment period or even install permanent measurement equipment. We assume that most environmental factors change slowly over time, so only a few measurements will be necessary in the first few years.

From the environmental measurement perspective, the Center Model has the advantage that it will be in a small number of geographic areas. It would therefore take less effort to carry out environmental measurements that are fairly constant over a large area, e.g., water quality if a single water supply system covers most of the area, some types of air pollutants, etc. The large sample sizes in the small number of centers also make the organization of the environmental data collection program simpler. However, the addresses of the sample women will not cluster to any great extent, so that measurement of local and small area environmental factors will require a fair amount of travel.

The sample in the Office Model will be in a sizable number of PSUs, and thus environmental measurements will face the same problems in the Household Model. In fact, the situation will be more costly; since there will be very little clustering within PSU, there may well be close to 118,000 different neighborhoods for which observations are necessary. From the point of view of environmental measures, the only advantage of the Office Model is that the recruitment, and hence the environmental measurements, could be done in a more compact time period.

5.2.4 Collection of Tissues in the Delivery Room

Information received late in the preparation of this report indicates that it will be important to collect placentas and umbilical cord blood samples at the time of birth. Obviously, the collection of placentas and cord blood samples will require the cooperation of both hospitals and obstetricians. This will undoubtedly be quite difficult and require considerable planning and substantial resources to accomplish successfully under any of the sampling models.

To some extent, each of the sampling models will face somewhat the same problems, namely, identifying the many hospitals at which the births will take place, deciding how to staff the hospitals and arrange for the proper resources at the required times, obtaining IRB approvals from each of the hospitals and from the patient, and identifying, storing, and shipping the samples. In addition, each of the models will face problems unique to its design. For example, in the Household Model, the dispersion of the sample will require dealing with a substantial number of hospitals and doctors. The United States contains about 5000 community hospitals, or an average of about 1.7 per county, but that undoubtedly varies strongly with county size, and large counties will be substantially represented in any of our designs. Thus, we estimate that the 800-county design will contain about 3.4 hospitals per county, the 300-county design will have about 5 hospitals per county, and a 100 county design will have about 7 hospitals per county. Overall, then, the number of hospitals in which births could take place would range from 700 to 2700, and this still would exclude those births occurring outside of these hospitals. If placentas and cord blood are essential for all mothers in the NCS, then modifications to the sample design for the Household Model may be required, e.g., selecting a smaller number of counties and/or selecting only parts of counties in counties with many hospitals. However, such modifications will likely weaken analyses that do not require the data from the placentas and cord blood.

The Office Model might be somewhat more manageable in that births might be more tightly clustered in a smaller number of hospitals by the use of doctors as the sampling frame. But, here too, the number of PSUs remains a major factor. As for the Center Model, it offers the likelihood of having to deal with the smallest number of hospitals, albeit still with all the other problems noted above.

One possibility to consider is whether the study requirements could still be met by selecting a subsample of mothers in a subsample of PSUs, and collecting the placenta and fetal material only from this group, which then would substantially limit both the number of hospitals and other birthing places and the amount of effort required. What is clear, however, is that, irrespective of the sample design, the collection of these materials will require substantial planning, a great deal of negotiation with doctors, hospitals, and mothers, and considerable resources of time and funding.

5.2.5 Long-Term Considerations

As the infants are born and grow into young adults, there will obviously be many moves—some of them over long distances. Statistics are regularly gathered in the Current Population Survey (CPS) on one-year mobility and population censuses have obtained data on five-year mobility.

More young children move over the course of a year than older persons. As many as 23 percent of children aged 1 to 4 in March of 2000 were living in a different home than in March of 1999 (Schachter, 2001). Annual mobility rates fall as the child ages (18% for those aged 5 to 9 and 15% for those aged 10 to 19) until reaching age 20, when they zoom up to the highest rate experienced during an individual's lifetime, 35 percent. Amongst the 23 percent of children aged 1 to 4 who moved, 61 percent had remained within the same county and 79 percent had remained within the same state. Over a five-year period, around 44 percent of the total population moves. It is thus clear that a sizable proportion of children will have moved from the county of their birth by the time that they reach the age of 20.

Regarding the importance of following children who move, a study of newborns in one Minnesota county (Katusic, et al, 1998) found that cumulative probability of migration out of the community by age 5 years to be 32.2% (95% confidence interval, 31.2 to 33.2%). This study found that the migrants had higher rates of congenital defect noted at birth (1.5% versus 0.7%). When considered simultaneously in a logistic regression model, the parents of migrants were more highly educated, migrant mothers were younger and had fewer prenatal visits, and migrant children were more likely to be black.

The Household and Office Models are clearly more capable of handling movers than the Center Model. This is because the Household and Office Models are much less clustered to begin with. If a highly distributed measurement system can be developed for the short term, it will stand up fairly well over the long term since a significant percentage of the movers will be within commuting distance of one of the sampled PSUs.

As one possible plan, suppose that a registered nurse is recruited in every PSU who can perform a wide variety of tasks—everything from drawing blood and collecting other biologics and anthropometric measurements, to administering questionnaires, including collecting data on maternal depression, family functioning and neighborhood characteristics. If it was possible to develop a system of measurement protocols that this one highly skilled nurse could conduct with her/his own home as a base of operations, then a plan with 800 PSUs might stand up very nicely over time. Similarly, having one recruited general-purpose clinic per PSU would also probably be an efficient system for obtaining physical measurements.

Mobility poses a significant problem for the environmental measurements in the longer term. A major effort would be required to collect information that uniquely reflects the environmental exposure of each individual over the 20-year life of the survey, other than what can be gleaned from samples collected by the parents of the sample youths or by ordinary survey interviewers who visit the children's homes. This is probably true regardless of the sample design.

Finally, we note a point made by a member of an expert panel who reviewed a draft of this report. Over 20 years, if the Center Model is used, some "center deaths" can be expected when a center lacks the resources or motivation to continue. Consideration of this issue is needed if the Center Model is adopted.

5.3 Hybrid Designs and Organization Structures

Reviewers of the first draft suggested various "hybrid" designs. The main point of these hybrids[32] was to try to combine the benefits of the Household Model for sampling such as generalizability and recruitment early in pregnancy with the benefits of the Center Model for the collection of placentas and other measurements that must be made in hospitals. This can be done, but probably only by sharply limiting the number of PSUs since the number of qualified medical centers is unlikely to be much greater than 100 and each center would probably only be able to supervise local operations. As we have noted elsewhere, there are both significant analytic advantages and operational problems attached to the use of a large number of PSUs to locate pregnant women. A hybrid design would require giving up those analytic advantages. The ultimate decision on the number of PSUs will depend on balancing these two considerations.

A secondary point of the hybrid designs appeared to be a preference for decentralized control of the study. Little has been said up to this point on the administrative structure for sample recruitment and all the subsequent measurements. A centralized hierarchical structure for all operations should result in the most uniform measurements. However, a decentralized structure might be able to better assemble and utilize a variety of resources as well as being more responsive to local issues. Each of the major types of data collection could be assigned to different agents in a centralized hierarchical structure. It would also be possible to use a centralized hierarchical structure for the recruitment phase while using a decentralized structure for all the other measurements provided that the number of PSUs was small enough, as indicated above.

[32] One of these was essentially the Household Model with 50 PSUs instead of 300 or 800. Another was to use the Center Model but encourage the centers to employ household screening. This, too, is very similar to the Household Model, but with a decentralized organizational structure.

6. COST PROJECTIONS FOR SAMPLE RECRUITMENT AND BASELINE INTERVIEWS

This chapter provides the cost detail for recruiting pregnant women and for conducting baseline interviews. As noted several times throughout this report, the costs of collecting, managing, storing, and analyzing biologics, physical exams, and environmental measurements are not included; nor are costs for long-term monitoring of the children included. Methods for the screening and recruiting phases may change substantially when considered as part of the overall data collection system rather than being considered in isolation and that may have a marked effect on costs. As plans for the other measurement systems are firmed up, we therefore recommend a comprehensive reexamination of the assumptions underlying the screening and recruiting phases to ensure effective integration of the operations and minimization of total cost.

Prior to presenting and discussing the cost estimates for the recruitment and baseline interviews, we describe more fully the key components of each design that were costed. The selection of which components should be included in the cost for each design was based on the rule that the cost estimate should be adequate to fund enrollment and administration of baseline interviews. As is noted further below, some components have uses beyond the initial enrollment and administration of baseline interviews. No attempt was made to prorate the cost of these dual-use components. Instead, if a component was required for the enrollment and baseline interview, then the full cost of that component was included.

In coming up with hours for specific tasks, many other assumptions were implicitly made about the required levels of review, types of software, systems requirements, types and levels of quality control, and other study requirements. Projected costs depend rather strongly on this full set of assumptions. In addition, and among other things, changes in government regulations for contractors or the issuance of special requirements for the NCS could materially affect the cost estimates.

When evaluating the value that is offered by a design, it is worth noting that even when different designs have a shared component, the quality of that component may not be the same across the designs. Since the Household, Office, and Center Models are so different from one another, it is not meaningful to enforce the same assumptions and modes of data collection on each of them. Rather, the approach adopted has been to develop separate approaches for each, constructed to best serve the particular model. For example, it cannot be assumed that the sample selection process for pregnant women in sampled offices in the Office Model is as well controlled as the selection of pregnant women in the Household Model. Issues of survey quality need to be taken into account in making cost comparisons between the models.

6.1 Components of Cost for the Household Model Designs

As discussed earlier, two sample designs from the household model were evaluated. They differ only in the numbers of PSUs and segments. Household Option A has 800 PSUs and 12,500 segments, while Household Option B has 300 PSUs and 3,125 segments. Here are the features they share:

- Development of a project management plan.
- A listing of all dwelling units in residential structures within the sample segments.
- A limited public relations campaign that includes primarily the development of a press release and distributing it within the sample PSUs to county executives, police chiefs, mayors, local newspapers, and local medical societies, as well as manning a hot line for questions and answers, but no advertising, no television material, and no local interviews of the project management by the local press.
- A set of regional supervisors who would be available for meetings with local officials in addition to normal work of supervising interviewers.
- Personal visits to all listed dwellings (1.3 million) to determine the presence of age eligible women.
- The use of proxy information from neighbors to determine age and gender eligibility of sample dwellings after some number of failed attempts to reach the occupants.
- An age eligibility window of 12 to 44 years (delayed eligibility for girls ages 12 to 14, assuming a three year screening cycle).
- A seven-day national training session for the interviewers, with repeat sessions every six months for new hires.
- A ten minute CAPI instrument to be used to screen some 500,000 women ages 15 to 44 for current pregnancy status and surgical sterility and to recruit those not currently pregnant or surgically sterile into a pregnancy monitoring system (all costs of questionnaire development including content development, programming, and testing).
- A restriction that the pregnancy screening instrument must be administered to the sample woman with no proxy response from other household members or anyone else.
- Permission for interviewers to conduct the pregnancy screener by phone for second and additional age-eligible women after the pregnancy screener is completed in person for one age-eligible woman in the household.

- Obtain signed parental consent for pregnancy screening interview for girls ages 15 to 17, unless married or otherwise living independently of parental figures.

- Establish an 800 number to receive reports of new pregnancies where incoming calls are automatically rolled over to interviewers who are working on active reminders to the age-eligible women.

- Telephone recontacts of age-eligible women who are not surgically sterile at three-month intervals to check for new pregnancies.

- Personal visits to recontact age eligible women who lack phone service or do not respond to the telephone recontact attempts.

- An $10 incentive payment to women who complete the pregnancy screener.

- A 45-minute complex Audi-CASI interview of all recruited pregnant women, including questions on such topics as substance abuse, occupation, social economic class, family structure, and diet (all costs of questionnaire development including content negotiation, programming, and testing).

- A $50 incentive payment to all pregnant women who complete the baseline interview.

- An electronic case management system that facilitates the supervision of the interviewers, keeps track of the disposition of all sample dwellings and women, and quickly moves the data to home-office computers.

- A help desk to respond quickly to interviewer problems with either the household or pregnancy screener or with the baseline interview.

- An appropriate level of security and redundancy in the data collection, processing, and management systems to protect the confidentiality of the data and preserve them from accidental loss and other security threats.

- Spanish versions of key questionnaires.

- Intensive testing of automated instruments.

- A pilot test of 2,400 addresses in three PSUs, with two months of screening, a three-month pause, and one month of telephone followup (should yield about 940 pregnancy screenings with age-eligible women and 54 baseline interviews with pregnant women).

- 10 percent of cases will be validated (primarily by phone but by personal visit where necessary).

- Clerical review and editing.

- Sampling weights and design-based variance estimation codes.

- Deliverable codebooks, archival respondent ID files, and archival analytic files.

- Monthly client management meetings and progress reports.

- Preparation and delivery of study data on a scheduled ongoing basis to facilitate additional data collection and quality control as well as serving analytic needs.

- A final methods report.

6.2 Components of Cost for the Office Model Design

As discussed earlier, one sample design from the Office Model was evaluated. It has 800 PSUs and 4000 offices. Here are other critical features that were assumed in the preparation of costs:

- Development of a project management plan.

- A sample of 40,800 MDs and DOs in targeted specialties in 800 PSUs (In many PSUs, this will mean taking a census of the MDs and DOs in the targeted specialties).

- A grouping operation where the doctors are sorted into joint practices based on address and insurance company data.

- A prescreening instrument for the offices to determine all of the practice locations for all the sample doctors and the volumes of pregnant women each doctor sees at each location.

- Prescreening of offices done by high level telephone interviewers. Operations last six months.

- A subsample of 5,700 practice locations for specific doctors—sampled in proportion to volume so that locations where doctors see many pregnant women are oversampled.

- A comprehensive IRB package to be distributed to all sampled doctors as part of the recruitment of the doctor as well as a tool for the doctor to use in requesting approval from any pertinent IRB.

- Visits by executive interviewers to all 5,700 doctors' office to recruit a sample of 4,000, and to provide the recruited offices with training on procedures. Operation lasts six months.

- A set of specific weeks during which each doctor (or their staff) is to induct all pregnant women he or she sees who are coming in for the first prenatal care visit following a confirmation of pregnancy (doctors with low volumes of pregnant women will have a longer set of weeks for recruitment; average number to be recruited per office is 32).

- A procedure for recruiting pregnant women where the doctor makes the first reference to the study, but where the bulk of the recruiting work is done by either a professional interviewer temporarily located in the office or by a nurse on the doctor's regular staff. Assume that 25 percent of offices have a professional recruiter and 75 percent have recruiting conducted by a nurse on the regular staff. Operations last 18 months, overlapping considerably with doctor recruiting, so that the elapsed time between the recruitment of the first and last pregnant women should be between 17 and 18 months.

- An average compensation package for sample doctors, their employers, or their practices of $150 for every pregnant woman recruited (averages to $4,800 per doctor).
- An average incentive to nurses when they do the recruiting of $50 per recruited pregnant woman (averages to $1,600 per nurse).
- A web-based application where doctors' staff enter the names, addresses, and phone numbers for recruited pregnant women into the computer systems of the institution managing the project.
- A help desk for doctors to call when there are problems (we estimated an average of 10 hours of support per office).
- A backup facility to receive faxed contact information for the recruited pregnant women from the doctors' offices should an office experience computer difficulties.
- A national field force of 300 interviews to conduct the Audio-CASI baseline interviews with the recruited women. Travel will be required given that there are 800 sample PSUs. It is assumed that they will achieve a 95 percent response rate. Operations last 19 months (equal to the time period for recruiting plus one month).
- A $50 incentive to all pregnant women who complete the baseline interview.
- A seven-day national training session for the interviewers, with repeat sessions every six months for new hires.
- A 45-minute complex Audi-CASI interview of all recruited pregnant women, including questions on such topics as substance abuse, occupation, social economic class, family structure, and diet (all costs of questionnaire development including content negotiation, programming, and testing), all administered by a local interviewer who calls on the women recruited in their doctors' offices.
- A help desk to respond to interviewer problems with the baseline questionnaire.
- An electronic case management system that facilitates the supervision of the interviewers, monitors the progress of the offices in meeting recruitment goals, keeps track of the disposition of all sample pregnant women, and quickly moves the data to home-office computers.
- An appropriate level of security and redundancy in the data collection, processing, and management systems to protect the confidentiality of the data and preserve them from accidental loss and other security threats.
- A Spanish version of the baseline questionnaire.
- Intensive testing of automated instruments.

- A pilot test of 75 doctors in three PSUs (should yield about 3,000 baseline interviews with pregnant women).
- 10 percent of cases will be validated (primarily by phone but by personal visit where necessary).
- Clerical review and editing.
- Sampling weights and design-based variance estimation codes.
- Deliverable codebooks, archival respondent ID files, and archival analytic files.
- Monthly client management meetings and progress reports.
- Preparation and delivery of study data on a scheduled ongoing basis to facilitate additional data collection and quality control as well as serving analytic needs.
- A final methods report.

6.3 Components of Cost for the Center Model Design

Separating the costs of recruitment and followup is natural and easy for the Household and Office Models, but somewhat misleading for the Center Model. The staff and physical assets employed in the Center Model for recruitment would stay on after recruitment and their roles would, in fact, evolve as the survey progresses. Depending on the measurements to be conducted, this might also be partly true for the home interviewers in Household and Office Models (for example, if normal interviewers are used to collect hair and urine specimens and to conduct followup interviews). Note, however, that this would not be true for the doctors in the Office Model since it would not desirable to have a group of 4,000 doctors who are mostly gynecologists conducting pediatric followups for 20 years.

One feature of the Center Model is that there is considerable paper work involved in awarding contracts to the 100 centers. This might be done either by a government agency or by a contractor. We have costed it out under the assumption that the work is done by a contractor, although we suspect that costs would be similar if the work was done by a government agency.

Given that preamble, here are the elements included in the cost estimate for the Center Model:

- Development of a project management plan.
- Proposals will be solicited by placing announcements in professional journals.

- The center-recruiting institution (either a government agency or a contractor) will host six half-day bidders' conferences.

- The center-recruiting institution will develop the RFP and post it on the web.

- The center-recruiting institution will travel to 30 particularly desired sites to convince them to submit bids (two days for two people).

- Three people at the center-recruiting institution will review, discuss, and rate each proposal.

- Center contracts will be fixed price.

- The indirect costs for the centers are assumed to have the same structure as that assumed for the center-recruiting institution. In addition, it is assumed that the center-recruiting institution will levy a 2.6 percent indirect charge on the center costs.

- A comprehensive IRB package to be distributed to all Centers would be prepared by a coordinating center or prime contractor.

- Each center will have a single physical location with space appropriate for ambulatory care (2,500 square feet, leased for three years at $20 per square foot per annum), a part-time medical director (15-25%), a half-time research nurse, and a full-time research assistant.

- It is assumed that many of the centers will not have adequate patient flow to recruit 1,200 women. Even those who do have adequate flows may want to make outreach efforts in order to fill quotas for pregnant women from different backgrounds. So beyond recruiting from their normal patient flows, it is assumed that all of the centers will use a small amount of advertising ($1,500 per month for 36 months comes to $54,000 per center before indirect charges). In addition, it is assumed that 60 percent will use referrals from a network of local doctors who receive $50 for each referral.

- Training for the research nurse and research assistant will be conducted by the coordinating center or prime contractor in a central location.

- A 45-minute complex Audi-CASI interview of all recruited pregnant women, including questions on such topics as substance abuse, occupation, social economic class, family structure, and diet (all costs of questionnaire development including content negotiation, programming, and testing) prepared by the Prime Contractor and administered in the Centers, with assistance to respondents provided to respondents by Center staff where required.

- Each recruited pregnant woman will receive a $50 incentive payment for answering the baseline questionnaire.

- A help desk to respond quickly to center problems with the baseline questionnaire.

- An appropriate level of security and redundancy in the data collection, processing, and management systems to protect the confidentiality of the data and preserve them from accidental loss and other security threats.

- A Spanish version of the baseline questionnaire.
- Intensive testing of automated instruments.
- Clerical review and editing.
- Deliverable codebooks, archival respondent ID files, and archival analytic files.
- Monthly client management meetings and progress reports.
- Preparation and delivery of study data on a scheduled ongoing basis to facilitate additional data collection and quality control as well as serving analytic needs.
- A final methods report.

6.4 Integration of Data Management Systems

The NCS will probably require a data management system that integrates control of the sampling and baseline interviews with control of the collection of biologics, information from physical exams, environmental measurements, etc. Our estimates include the cost for a simpler data management system—one that will only control the sampling and baseline interviews. The data management system for control of data on health outcomes and environmental exposure will be considerably more complex than the system required for recruitment. If enough is known about the requirements for the more complex data management at the time that the simpler system is written, there may be some cost efficiencies, but it would be safer to assume that the costs for the two data management systems would be additive.

6.5 Cost Estimates for Recruitment and Baseline Interviews

Table 6-1 compares the costs of the four designs. Labor hours are translated into millions of 2002 dollars. No attempt has been made to predict inflation from the present until the conclusion of recruitment. Other direct costs include items such as computer leases, telephone charges, travel expenses, and so on. Costs have been loaded with indirect costs and fees that are within the range charged by various contractors but do not reflect those of any specific contractor.

Table 6-1. Costs for recruitment and baseline interviews for each of the four designs

Component	Household Model A	Household Model B	Office Model	Center Model
Labor Hours (1,000s)				
Project Management	26	26	15	18
Field Management	847	541	244	66
Statistical Operations	10	10	9	4
Systems Programming	319	327	211	100
Other Professional Staff	13	13	13	115
Support Staff	193	244	98	42
Interviewers	3,591	3,040	1,111	0
Millions of 2002 Dollars (including estimated indirect costs and fees)				
Labor Costs	$139	$121	$59	$27
Payments to Medical Centers and Providers	$0	$0	$19	$130
Respondent Incentives	$11	$11	$6	$6
Other Costs	$39	$31	$26	$3
Total	$189	$163	$109	$166

The Household Model with 800 PSUs is the most expensive. Household Option A is projected to cost $189 million. With the cost savings from greater clustering, Household Option B is projected to cost $163 million. Note however, that power levels are much lower with Option B than with Option A, as discussed in Chapter 4. The Center Model has about the same cost as Household Option B with a projected cost of $166 million. The Office Model is considerably less expensive at $109.

The Office Model has much lower cost for recruitment and baseline interviews than the Household Model options because there is no attempt to recruit women who never seek prenatal care nor to recruit them as soon as they think they are pregnant. Costs would be some $30 million lower in the Office Model if the baseline questionnaire could be administered in the doctors' offices with assistance from office staff, rather than sending professional interviewers to the women's homes. However, this added burden on physicians might lead to serious additional physician recruitment failures. Costs for environmental measurements under the Office Model may be much higher than those for the Household Model given the dispersion of the sample, while costs of collection of biologics from the pregnant women would be much lower in the Office model if done by office staff, but quantifying these considerations is beyond our scope.

The Office Model has a much lower cost for recruitment and baseline interviews than the Center Model because the Office Model takes advantage of the existing infrastructure and the natural flow of pregnant women to physicians. The Center Model requires creating some new infrastructure (space, a skeleton staff, and an organizational charter) and often taking some measures to generate an adequate traffic of pregnant women. Of course, the infrastructure will be useful for health outcomes measurements and those exposure measurements that involve analysis of biologics and tissues for the duration of the NCS. As an example of this, the administration of the baseline questionnaire is far less expensive in the Center Model than in the Office model because we assume that the interviews are conducted at the centers with assistance from center staff, whereas we assume that interviews are conducted at women's homes with assistance from professional survey interviewers with the Office Model.

Comparing the Household and Center Models, the costs for recruitment and baseline interviews are very similar if the more clustered Household Design is selected. In terms of costs for health outcomes and environmental exposure measurements, the Center Model is likely to be considerably less costly for health outcomes and the exposures measured through biologics and tissues, while the Household Model is likely to be less costly for those exposures measured through direct collection of samples from the home environment. Again, trying to determine the extent to which these other costs might offset or magnify the initial cost difference is beyond our scope.

7. OVERALL SUMMARY

This chapter summarizes the properties of each design with respect to bias, power, cost, and other features. In this brief review, important subtleties are, of course, lost. We urge a thorough review of the entire report to gain a nuanced perspective on the properties of the designs. One point that needs to be emphasized here is that, had more been known about the measurement protocols for health outcomes and exposures, we believe it likely that we would have crafted somewhat different designs. As a result, we recommend a thorough re-examination of the designs once the measurement protocols have been established. With that caveat, we now list positive features of each of the four designs, starting with Household Models A and B.

Positive features of Household Models A (800 PSUs & 12,500 segments) and B (300 PSUs and 3,125 segments)

+ A probability sample of virtually all pregnant women
+ Coverage of women who never seek prenatal care
+ The ability to recruit women early in their pregnancy
+ A better response rate than the Office Model
+ Very good statistical power for Model A; reasonable power for Model B
+ The ability to use survey interviewers to gather environmental samples from women's homes if appropriate
+ A rich sample of miscarriages for study
+ Can be adapted to recruit women prior to pregnancy (although at additional cost)
+ Excellent geographic dispersion
+ Strong control over recruitment operations
+ No need to deal directly with obstetricians, local IRBs, or local research directors to recruit the sample
+ Household Model B has lower costs than Household Model A, particularly for segment-level measurements (if required)

Positive features of the Office Model

+ A probability sample of pregnant women seeking prenatal care
+ Very good statistical power
+ The ability to use survey interviewers to select environmental samples from women's homes if appropriate
+ Excellent geographic dispersion
+ Low recruitment cost
+ Low cost for conducting maternal exams and collecting maternal biologics (if done in the offices)
+ Ability to recruit the sample quickly
+ No need to deal with local research directors to recruit the sample

Positive features of the Center Model

- + Low recruitment cost if the cost of setting up the centers is excluded
- + Low cost for conducting maternal exams and collecting maternal biologics
- + Most feasible design for collection of placentas and cord blood samples
- + The small number of centers facilitates the ability to thoroughly train the centers' staffs in measurement protocols in order to give standardized health measurements
- + Excellent control over recruiting operations
- + Similar models to recruit pregnant women have been used successfully in previous studies

Chapter 6 presents cost estimates for recruiting the sample of pregnant women and conducting baseline interviews with them under all four models. The total cost estimates are: $189 million for Household Model A, $163 million for Household Model B, $109 million for the Office Model, and $166 million for the Center Model. However, we caution strongly against simple comparisons of these costs for two main reasons. First, given the very different natures of the four designs, the data collection methodologies are correspondingly very different, and hence the products are also different. Second, if enough information on measurement protocols were available to make the 25-year cost for each design estimable, the relative costs of the four designs might be completely rearranged. For example, it seems likely that the 25-year cost of the Office Model will be higher than the 25-year cost of the Center Model even though we estimate that it will take $57 million more to launch the Center Model than to launch the Office Model.

Another consideration about costs is that, even though the range of the estimated costs for recruitment and baseline interviews across the models is about $880 million, this is still a fairly modest portion of the projected total $2 billion cost over 25 years. Further, the initial investment in a high-quality sample will pay dividends throughout the life of the study. Hence considerably higher recruitment costs may well be fully justified.

As a final point, we strongly urge that extensive pilot studies be conducted to examine the operational feasibility and re-estimate costs for all three models. In developing the models we have made a number of critical assumptions that need to be tested empirically. For example, with the Household Model, the effectiveness of a pregnancy screening instrument in identifying pregnant women needs to be tested, as does its ability to achieve high response rates. The Office Model assumes the participation of about half of sampled physicians: Can that or a higher rate be achieved, and what incentives are required? There is perhaps less need for experimentation with the Center Model since smaller scale precedents do exist. However, studies are required to assess, for example, its ability to recruit centers that cover all parts of the U.S. population. Such a large-scale and ambitious study as the NCS demands extensive pilot work before the full-scale study is launched.

REFERENCES

Bachrach, C.A., Horn, M.C., Mosher, W.D., and Shimizu, I. (1985). *National Survey of Family Growth, Cycle III, Sample Design, Weighting, and Variance Estimation*. Series 2 of Vital and Health Statistics, No. 98. DHHS Pub No. (PHS) 85-1372. Hyattsville, MD. National Center for Health Statistics.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hiearchical Linear Models: Applications and Data Analysis Methods.* Newbury Park, CA: Sage.

Chandra, A. and Stephen, E. H. (1998). Impaired Fecundity in the United States: 1982-1995. Family Planning Perspectives, 30, 34-42.

Chromy, J.R., Bowman, K.R., Crump, C.J., Packer, L.E., and Penne, M.A. (1999). Population Coverage in the National Household Survey on Drug Abuse. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 576-580. Alexandria, VA: American Statistical Association.

Cornfield, J. (1978). Randomization by group, a formal analysis. *American Journal of Epidemiology*, 108, 100-102.

Cunningham FG, Gant NF, Leveno KJ, Gilstrap LC, Hauth JC, Wenstrom KD. (2001). Williams Obstetrics, 21st edition. New York, McGraw-Hill.

Donner, A. and Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research.* London: Arnold.

Goldstein, H. (1995). *Multilevel Statistical Models*, 2nd ed. London: Edward Arnold.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*. New York: John Wiley.

Hox, J. (2002). *Multilevel Analysis: Techniques and Applications*. Mawah, NJ: Lawrence Erlbaum Associates.

Judkins, D.R., Chu, A., DiGaetano, R., and Shapiro, G. (1999). Coverage in Screening Surveys at Westat. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 581-586. Alexandria, VA: American Statistical Association.

Judkins, D.R, Shapiro, G., Brick, M., Flores-Cervantes, I., Ferraro, D., Strickler, T., and Waksberg, J. (1999). *1997 NSAF Sample Design, in NSAF Methodology Reports*. Washington, DC: Urban Institute.

Kalsbeek, W.D., and Macewicz, M.J. (1993). Sampling prenatal care providers from a frame of physicians. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 206-211. Alexandria, VA: American Statistical Association.

Kalton, G. and Blunden, R. M. (1973). Sampling errors in the British General Household Survey. Bulletin of the International Statistical Institute, Book 3 of the 1973 Proceedings, pp 83-97.

Katusic SK, Colligan RC, Barbaresi WJ, Schaid DJ, Jacobsen SJ. (1998). Potential influence of migration bias in birth cohort studies. *Mayo Clin Proc* ,73, 1053-61.

Kelly, J.E., Mosher, W.D., Duffer, A.P., and Kinsey, S.H. (1997). *Plan and Operation of the 1995 National Survey of Family Growth*. Series 1 of Vital and Health Statistics, No. 36. DHHS Pub No. (PHS) 98-1313. Hyattsville, MD. National Center for Health Statistics.

Khare, M., Mohadjer, L.K., Ezzati-Rice, T.A., and Waksberg, J. (1994). An evaluation of nonresponse bias in NHANES III (1998-1991). *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 949-954. Alexandria, VA: American Statistical Association.

Kish, L. and Frankel, M. F. (1974). Inference from Complex Samples (with discussion). *Journal of the Royal Statistical Society, Series B,* 36, 1-37.

Lê, T.N. and Verma, V.K. (1997). *An Analysis of Sample Designs and Sampling Errors of the Demographic and Health Surveys*. Demographic and Health Surveys, Analytic Reports No. 3. Calverton, MD. Macro International.

Longford, N. T. (1993). *Random Coefficient Models.* Oxford: Clarendon Press.

Mohadjer, L.K, Montaquila, J., Waksberg, J, Bell, B., James, P., Flores-Cervantes, I., and Montes, M. (1996). *National Health and Nutrition Examination Survey III Weighting and Estimation Methodology.* Rockville: Westat.

National Center for Health Statistics. (1973). *Reliability of Estimates With Alternate Cluster Sizes in the Health Interview Survey*. Series 2 of Vital and Health Statistics, No. 52. DHEW Pub No. (HSM) 73-1326. Hyattsville, MD. National Center for Health Statistics.

Niswander, K.R. and Gordon, M. (1972). *The Women and Their Pregnancies.* Washington, DC: U.S. Government Printing Office.

Olsen J, Melbye M, Olsen SF, Sorensen TI, Aaby P, Andersen AM, Taxbol D, Hansen KD, Juhl M, Schow TB, Sorensen HT, Andresen J, Mortensen EL, Olesen AW, Sondergaard C. (2001). The Danish National Birth Cohort-- its background, structure and aim. *Scand J Public Health*. 29, 300-7.

Power, C., Hertzman, C., and Matthews, S (1997). Social differences in health: life-cycle effects between ages 23 and 33 in the 1958 British Birth Cohort. *American Journal Of Public Health*, 87, 1499-503.

Sanders, L.L. and Kalsbeek, W.D. (1990). Network sampling as an approach to sampling pregnant women. *Proceedings of the Section on Survey Research Methods of the American Statistical Association*, pp. 326-331. Alexandria, VA: American Statistical Association.

Schachter, J. (2001). *Geographic Mobility: Population Characteristics*. Current Population Reports, pp. 20-538. Suitland, MD: U.S. Census Bureau.

Wadsworth, M. E. J., Mann, S. L., Rodgers, B. (1992). Loss and representativeness in a 43 year follow up of a national birth cohort. *Journal of Epidemiology and Community Health*. 46, 300-304.

Waksberg, J., Judkins, D., and Massey, J. T. (1997). Geographic-based oversampling in demographic surveys of the United States. *Survey Methodology*, 23, 61-71.

Waksberg, J., Sperry, S., Judkins, D.R., and Smith, V. (1993). *National Survey of Family Growth, Evaluation of Linked Design.* (Vital Health Statistics 2 (117), (PHS) 93-1391). Hyattsville, MD. National Center for Health Statistics.

Wolf, L. E., Croughan, M., and Lo, B. (2002). The challenges of IRB review and human subjects protections in practice-based research. *Medical Care*, 40, 521-529.

Woodward, L. J. and Fergusson, D. M. (2001). Life course outcomes of young people with anxiety disorders in adolescence. *Journal Of The American Academy Of Child And Adolescent Psychiatry*, 40, 1086-93.

Zornberg, G. L., Buka, S. L., & Tsuang, M. T (2000). Hypoxic-ischemia-related fetal/neonatal complications and risk of schizophrenia and other nonaffective psychoses: A 19-year longitudinal study. *American Journal Of Psychiatry*, 157, 196-202.